WHAT EMPIRICAL TRAINING DYNAMICS DO NOT DETERMINE: RELU HISTORIES AND ROBUST PREDICTION SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether a ReLU network's empirical training record determines its breakpoint history, the order in which its breakpoints cross during training. The record includes training predictions, loss, the empirical neural tangent kernel, parameter and gradient norms, and the Hessian spectrum. For shallow one-dimensional networks with two training samples, we find an open set of initial parameters. From each point in this set, every stretchable complete reversal can occur from another initialization with the same record on a common finite gradient-flow interval. We also give a common gradient-descent schedule. A three-sample construction gives a prediction gap above 0.04 on a fixed interval after 16 steps at learning rate 0.005. This gap survives prescribed data perturbations and independent initialization and update errors. Further examples use two-dimensional data and two hidden layers, including a training activation change. These are selected constructions, not claims about typical initialization or generalization.

1 INTRODUCTION

One may monitor the training of a neural network by its predictions on the training sample, and a set of standard diagnostics including the loss function, the empirical neural tangent kernel (NTK), parameter and gradient norms, and Hessian spectrum. We call the resulting time series the empirical training record.

Our question is whether this record determines how the combinatorics of the learned function changes between the training samples. This question matters when training statistics are used to compare two runs or to explain what a network has learned.

A shallow ReLU network with one-dimensional input and output represents a piecewise-linear function on $\mathbb{R}$. A breakpoint is a point where its slope changes. Breakpoints move during training and may cross one another. We label them by their initial left-to-right order. The ordered list of their crossings is the breakpoint history. We keep the relevant output weights and slopes positive, so the breakpoints are exactly the places where one of the ReLU inputs becomes 0.

Main results. **A large set of histories from the same empirical training record.** Our first result concerns shallow networks with one input, one output and two training samples. In an open set of initial parameters, each initialization can be matched with another that realizes any stretchable complete reversal and has the same empirical training record on a common finite gradient-flow interval. A corresponding result holds for a shared gradient-descent schedule chosen below a prescribed learning-rate cap (Theorem 3.1). The later examples extend the observed ambiguity, not this complete-history statement, to two-dimensional inputs and two hidden layers.

Further examples beyond the basic model.

The phenomenon also occurs beyond the two-sample setting. We give exact reference pairs with the same empirical record and different breakpoint histories in the following settings:

1. One hidden layer, one-dimensional input and output, and three training samples.
2. One hidden layer, two-dimensional input, scalar output, and four training samples.

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.

3. Two hidden layers, two-dimensional input, scalar output, and four training samples.

The three-sample pair also has a fixed-rate robustness guarantee. After 16 GD steps at rate 0.005, its prediction gap exceeds 0.04 on a fixed interval under the data, initialization and update budgets in Theorem J.2. The perturbed records satisfy the stated error bands; they need not be exactly equal.

Relation to prior work. Zhao et al. (2023) study loss-preserving symmetries, including nonlinear, data-dependent transformations that can change predictions at other inputs. Grigsby et al. (2023) study additional symmetries that preserve the full represented function, beyond neuron permutations and positive rescalings. We require more than equality at selected inputs: the original Euclidean optimizer must realize the specified breakpoint histories while preserving the full time-indexed empirical record. Related non-injectivity under the exact path kernel (EPK) representation is discussed by Bell et al. (2023, Section 3.6).

The dynamics of shallow one-dimensional ReLU networks have also been studied through reduced and spline representations (Williams et al., 2019; Sahs et al., 2022). Scale changes may preserve a function while changing its gradient-flow dynamics (Sahs et al., 2022, Section 2.5.1), and Nurisso et al. (2024) study invariant sets associated with individual neurons. We ask a different inverse question: which breakpoint histories can occur from one fixed empirical state under the original optimizer?

We use simple allowable-sequence realizability from Hoffmann and Merckx (2018) and the staircase-tableau/sorting-network correspondence and count of Araujo et al. (2026, Section 3 and Theorem 1.3). Under the standard point–line correspondence, these objects encode crossing orders produced by sweeping affine lines. Our contribution is to realize these orders with the original gradient flow and gradient descent while keeping the empirical record fixed. The equality of Hessian spectra follows from a fixed Euclidean isometry, consistent with earlier symmetry-based curvature analysis (Arjevani and Field, 2020). Recent work on local training geometry (Bona-Pellissier et al., 2026) and on open sets of parameters identifiable from the full function, up to permutations and positive rescalings, under specified architecture conditions (Grillo and Montúfar, 2026, Theorem 4.13) does not rule out different full functions sharing the restricted empirical records studied here.

2 Model, records, and histories

Let Θ be the set of parameters of a ReLU neural network for a given architecture. For every $\theta \in \Theta$, let f_θ be the corresponding piecewise-linear function. For a training dataset $d = \{(x_i, y_i, \lambda_i)\}_{i=0}^{n-1}$ with weights $\lambda_i > 0$, consider the squared loss

$$L_d(\theta) = \frac{1}{2} \sum_{i=0}^{n-1} \lambda_i (f_\theta(x_i) - y_i)^2. \tag{1}$$

We train all parameters in the original Euclidean metric. Gradient flow (GF) satisfies

$$\dot{\theta} = -\nabla L_d(\theta)$$

and by gradient descent (GD) we mean the sequence

$$\theta_{k+1} = \theta_k - \eta_k \nabla L_d(\theta_k)$$

Now consider the special case of ReLU neural networks with one input, one output, and one hidden layer with m neurons. Write $\theta = ((a_j, w_j, b_j)_{j=1}^m, c) \in \mathbb{R}^{3m+1}$,

$$f_\theta(x) = c + \sum_{j=1}^{m} a_j \,\mathrm{ReLU}(w_j x + b_j)$$

When $a_j, w_j > 0$, the breakpoints are $t_j = -b_j/w_j$. A history records the initial breakpoint order followed by the labels of each isolated crossing pair. GF crossings are transverse, with no simultaneous or endpoint events. For GD, we follow the line segments joining successive parameter vectors. An intermediate node crossing belongs to the preceding half-open step $(k, k+1]$ and must have the same crossing direction on both adjacent segments; a contact followed by a return is not a crossing.

We use the empirical neural tangent kernel (NTK), the NTK Gram matrix on the training inputs (Jacot et al., 2018, Section 4), and collect the following observations:

$$K_{\mathrm{emp}}(i,k;\theta) = \langle \nabla_\theta f_\theta(x_i), \nabla_\theta f_\theta(x_k) \rangle,$$
$$\mathcal{O}_d(\theta) = \big(f_\theta(X), L_d(\theta), K_{\mathrm{emp}}(\theta), \|\theta\|_2, \|\nabla L_d(\theta)\|_2, \mathrm{spec}^\uparrow \nabla^2 L_d(\theta)\big). \tag{2}$$

Here the spectra are sorted with multiplicity. Equality of spectra does not mean equality of the Hessian matrices in the original coordinates. Derivative-based records are evaluated where the derivatives exist; at a training activation boundary, we compare one-sided limits rather than assign a classical Hessian. Extra parameter readouts used in the experiments are not part of $\mathcal{O}_d$.

For higher-dimensional inputs, we use the same loss and record. We restrict the input to a fixed line and follow specified breakpoints along it. This slice history does not describe the full region-adjacency structure of the network.

3 TWO-SAMPLE CASE

A *stretchable complete reversal* is the order of the $\binom{m}{2}$ pairwise crossings in a sweep of affine lines $l_j(z) = x_j + y_j z$, with all crossing times distinct and in $(0,1)$. We label the lines by their initial top-to-bottom order, so $l_1(0) > \cdots > l_m(0)$ and $l_1(1) < \cdots < l_m(1)$. This matches the initial left-to-right breakpoint labels used below. The full chronological order is retained (Hoffmann and Merckx, 2018); see also Araujo et al. (2026, Section 3, especially Definition 3.1).

The first main result concerns shallow neural networks with one input and one output:

Theorem 3.1. *Consider a ReLU network with one input, one output and one hidden layer of $m \geq 2$ neurons,*

$$f_\theta(x) = c + \sum_{j=1}^{m} a_j \,\mathrm{ReLU}(w_j x + b_j),$$

and the training set $\{(0,0,1),(1,0,1)\}$. There is a nonempty open set Y in the region $c < 0$, $a_j, w_j > 0$, $b_j < 0$ and $-b_j/w_j < 1$ with the following property. For every $\theta \in Y$ and every stretchable complete reversal ω, there is $\theta_\omega \in Y$ whose original GF path realizes ω on a finite interval and has the same empirical training record as the path from θ on that interval. For GD, the same conclusion holds at matching nodes and interpolation fractions for a suitable common constant learning rate and finite step count. The rate can be chosen below any prescribed positive cap.

Remark 3.2. *Embedding the data and weights in one input coordinate gives an example with more input dimensions, but the added directions do not participate in training. Example 2 instead uses non-collinear data and trains both input directions. This embedding alone does not establish the same empirical-record statement for vector outputs.*

Remark 3.3. *Figure 1 illustrates the prescribed-state construction in Theorem C.2, which underlies Theorem 3.1. The two runs follow different stretchable complete reversals and have nearly coincident sampled records.*

Proof. Write $u_j = w_j + b_j$ and $q_j = w_j - b_j$. On the stated region, all hidden units are inactive at 0 and active at 1. The original equations are

$$\dot{a}_j = -f(1)u_j, \qquad \dot{w}_j = \dot{b}_j = -f(1)a_j, \qquad \dot{c} = -(c + f(1)).$$

Set $D = \sum_j a_j u_j$ and $E = \sum_j (u_j^2 + 2a_j^2)$. Then

$$\dot{c} = -(2c + D), \qquad \dot{D} = -(c+D)E, \qquad \dot{E} = -8(c+D)D.$$

These equations close the empirical dynamics. The parameter norm also depends on $\sum_j q_j^2$, since $\|\theta\|^2 = c^2 + \frac{1}{2}(E + \sum_j q_j^2)$. We therefore match this sum as well. For the given initialization, put

$$q_*^2 = \frac{1}{m}\sum_j q_j^2, \qquad (c_0, D_0, E_0) = (c, D, E).$$

Choose a centered line witness $x_j + y_j z$ for ω, with $|x_j|, |y_j| \le 1/2$, and write $\mu_x = m^{-1}\sum_j x_j^2$, $\mu_y = m^{-1}\sum_j y_j^2$ and $\mu_{xy} = m^{-1}\sum_j x_j y_j$. For a fixed sufficiently small $\epsilon > 0$, set

$$p = \frac{D_0}{mq_*^2} + \epsilon^3 \mu_{xy}, \qquad s = \frac{E_0}{mq_*^2} - \epsilon^4 \mu_x - 2\epsilon^2 \mu_y,$$
$$V = \sqrt{(s + \sqrt{s^2 - 8p^2})/4}, \quad H = p/V.$$

Take $h_j = H + \epsilon^2 x_j$, $v_j = V - \epsilon y_j$, and

$$(a_j, w_j, b_j) = \left(q_* v_j, \frac{q_*}{2}(1 + h_j), \frac{q_*}{2}(h_j - 1)\right), \qquad c = c_0.$$

Centering gives $D = D_0$, $E = E_0$ and $\sum_j q_j^2 = mq_*^2$ exactly. The resulting paths therefore have the same predictions, loss, kernel and norm records. A fixed orthogonal change in the mode coordinates of Appendix B, including the q coordinates, also gives the same Hessian spectra.

For GF, define $\tau(t) = \int_0^t (c(s) + D(s))\, ds$ and $\xi = \tanh(\sqrt{2}\tau)/\sqrt{2}$. Direct solution gives

$$\frac{w_j(t) + b_j(t)}{w_j(t) - b_j(t)} = \frac{h_j - 2v_j\xi(t)}{\sqrt{1 - 2\xi(t)^2}}, \qquad h_j - 2v_j\xi = H - 2V\xi + \epsilon^2(x_j + y_j\, 2\xi/\epsilon).$$

The common denominator is positive. Breakpoint order is the reverse of the order of u_j/q_j, so it follows the top-to-bottom line order. Appendix C.4 specifies the open set Y and proves that both paths reach $\xi = \epsilon/2$ before leaving the region. The GD argument uses the exact common discrete clock in Appendix C; approximation of GF alone is not used to infer exact equality of records. □

4 Further Examples

The exact reference pairs below have the same empirical record and different histories. Theorems J.2, O.1 and P.1 give their respective perturbation guarantees using analytic bounds and finite arithmetic certificates. Experiments in Section 6 check saved numerical runs separately.

4.1 Example 1

Consider the ReLU neural network with 1 input and 1 output, one hidden layer with three neurons. In other words,

$$f(x) = c + \sum_{j=1}^{3} a_j \,\mathrm{ReLU}(w_j x + b_j)$$

The training set is

$$\{(x_i, y_i, \lambda_i)\} = \{(0, 0, 1), (1, 0, 1), (2, 0, 1)\}$$

Define two rational matrices:

$$Z_A = \begin{pmatrix} 3/5 & 9/10 & -9/50 \\ 11/10 & 13/10 & -13/25 \\ 4/5 & 17/10 & -221/200 \end{pmatrix}, \quad Q = \frac{1}{27}\begin{pmatrix} 25 & 10 & 2 \\ -10 & 23 & 10 \\ 2 & -10 & 25 \end{pmatrix}, \quad Z_B = QZ_A, \qquad (3)$$

This construction is related to the linear loss-symmetry framework of Zhao et al. (2023, Section 3.3).

Now set the two parameters θ_A and θ_B as follows:

- $c_A = c_B = -1/2$.
- The three numbers on the j-th rows of matrix Z_A are the a_j, w_j and b_j, respectively, for θ_A.
- The three numbers on the j-th row of matrix Z_B are the a_j, w_j and b_j, respectively, for θ_B.

Train both networks by GD with $\eta = 1/200$ for $N = 16$ steps. Their exact empirical records agree. A has no breakpoint collision, while B has one collision in steps 8–15.

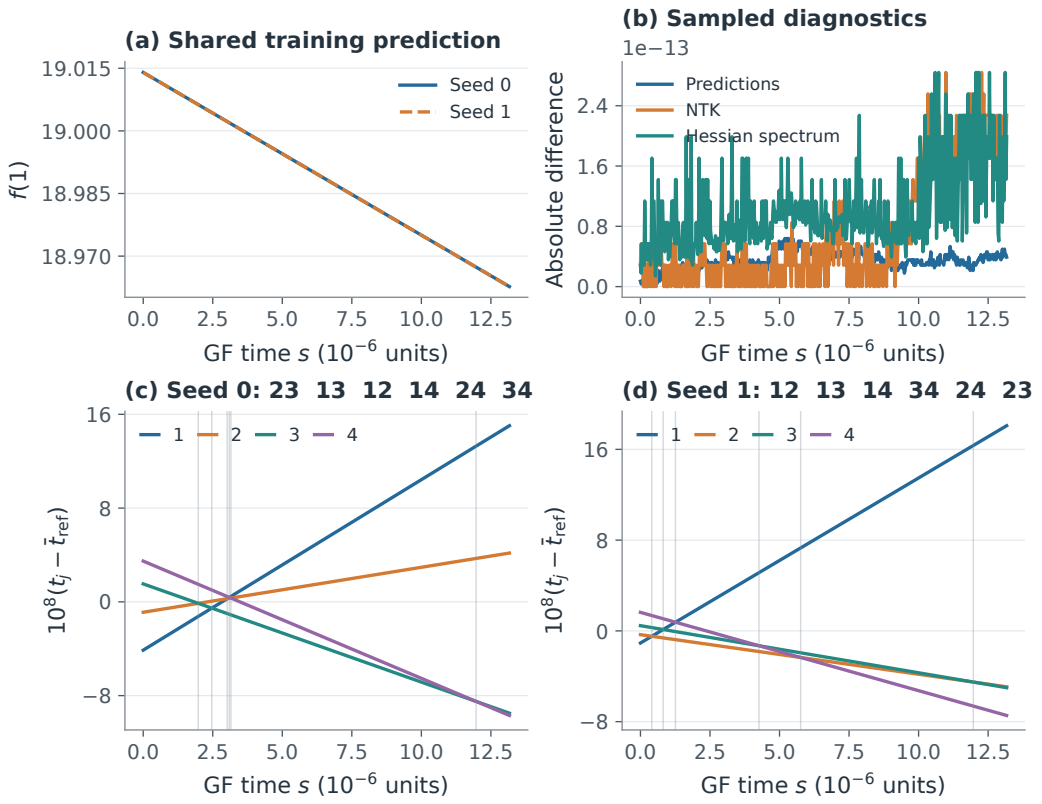


Figure 1: Nearly coincident empirical diagnostics accompany different breakpoint histories. These are the fixed E1 width-4 center witnesses, seeds $0/1$. Panels (c,d) subtract the same seed-0 mean $\bar{t}_{\mathrm{ref}}$ at each common GF time and scale by 10^8; their titles list the saved anchored collision words. The horizontal coordinate is $10^6 s$, with s the ODE time parameter, not wall-clock time. Vertical lines mark saved event times; visual crossings do not certify their accuracy.

4.2 Example 2

Consider the ReLU neural network with 2 input and 1 output, one hidden layer with 3 neurons. In other words,

$$f(x,y) = c + \sum_{j=1}^{3} a_j ReLU(w_{x,j}x + w_{y,j}y + b_j)$$

The training sample $\{(x_i, y_i, z_i, \lambda_i)\}$ equals

$$\{(0,0,0,1),(1,-1/2,0,1/2),(1,1/2,0,1/2),(2,0,0,1)\}$$

Now define the two rational matrices as:

$$Z_A = \begin{pmatrix} 3/5 & 9/10 & 1/10 & -9/50 \\ 11/10 & 13/10 & 3/20 & -13/25 \\ 4/5 & 17/10 & -1/20 & -221/200 \end{pmatrix}, \quad Q = \frac{1}{27}\begin{pmatrix} 25 & 10 & 2 \\ -10 & 23 & 10 \\ 2 & -10 & 25 \end{pmatrix}, \quad Z_B = QZ_A, \tag{4}$$

Set the two parameters θ_A and θ_B as follows:

- $c_A = c_B = -1/2$.
- The four entries on the j-th rows of matrix Z_A are the a_j, $w_{x,j}$, $w_{y,j}$, and b_j, respectively, for θ_A.

- The four entries on the j-th row of matrix Z_B are the a_j, $w_{x,j}$, $w_{y,j}$, and b_j, respectively, for θ_B.

Now do gradient descent with step size $\eta = 1/200$ and number of steps $N = 16$. Then, when restricted to the line $y = 0$, the trajectory started with θ_A has no breakpoint collision, while the trajectory started with θ_B has a breakpoint collision between steps 8 and 15.

4.3 Example 3

Consider a ReLU network with two inputs, one output, and hidden layers of widths 2 and 4:

$$u_\ell(x, y) = \mathrm{ReLU}(p_{\ell,x}x + p_{\ell,y}y + p_\ell), \qquad f(x, y) = c + \sum_{j=1}^{4} a_j \,\mathrm{ReLU}(b_j + v_{j,1}u_1 + v_{j,2}u_2).$$

Use the same four weighted training samples as in Example 2. Define

$$P = \begin{pmatrix} 1 & 1/5 & 1/10000 \\ 1/5 & 1 & 1 \end{pmatrix}, \qquad V_B = \begin{pmatrix} 1 & 1/20 \\ 13/10 & 1/10 \\ 9/10 & 3/20 \\ 1/10 & 1/5 \end{pmatrix},$$

$$a_B = (3/5, 1, 3/10, 1/5)^\top, \qquad s = (9/20, 5999/10000, 6001/10000).$$

Let p_x and p_b be the first and third columns of P. Set

$$(b_B)_k = -(V_B p_x)_k s_k - (V_B p_b)_k \quad (k = 1, 2, 3), \qquad (b_B)_4 = 1/5.$$

With the same Q as in Example 2, put $Q_4 = \mathrm{diag}(Q, 1)$ and

$$(V_A, b_A, a_A) = (Q_4^\top V_B, Q_4^\top b_B, Q_4^\top a_B).$$

Both initializations use this P and $c = -1/2$. Row ℓ of P gives $(p_{\ell,x}, p_{\ell,y}, p_\ell)$, and row j of V gives $(v_{j,1}, v_{j,2})$. All 23 parameters are trained by GD with $\eta = 1/50000$ and $N = 128$. The first lower unit deactivates at $(0, 0)$ in the first parameter segment. On $y = 0$, $x \in [1/4, 7/10]$, the three tracked upper-layer breakpoints have no collision in A and one $(2, 3)$ collision in B, in steps 2–4. The exact reference paths have the same empirical record on each smooth branch, with matching one-sided records at the training boundary. Appendix P gives the robustness statement for this configuration.

5 Further Remarks

Functions along the GF trajectory At a positive strict-order point, the function determines its breakpoint locations and slope jumps through the spline representation (Sahs et al., 2022, Section 2.1). However, without knowing the exact parameters, identical initial full functions and empirical traces can still have different later histories (Proposition I.3, under different hypotheses from Example 1 in Section 4).

Computational Complexity For supplied rational parameters in the setting of the two-sample main theorem, an algorithm can be used to detect attained GF breakpoint crossings from infinite-time limits and computes the maximal crossing sequences in polynomial bit complexity; it does not give event times at a prescribed physical horizon (Theorem E.1, Corollary E.2). Conversely, realization at the central empirical state is $\exists\mathbb{R}$-complete by simple allowable-sequence realizability (Hoffmann and Merckx, 2018, Theorem 6). For GD the rate below a cap and finite step count are existential, not preassigned or polynomially bounded (Theorem H.1).

6 Experiments

We report five numerical experiments, E1–E5, to illustrate the main theorem and the examples. E1–E3 retain the original protocol and results in Appendix M. E4 and E5 add 322 network pairs in a separate archive, with details in Appendix R.

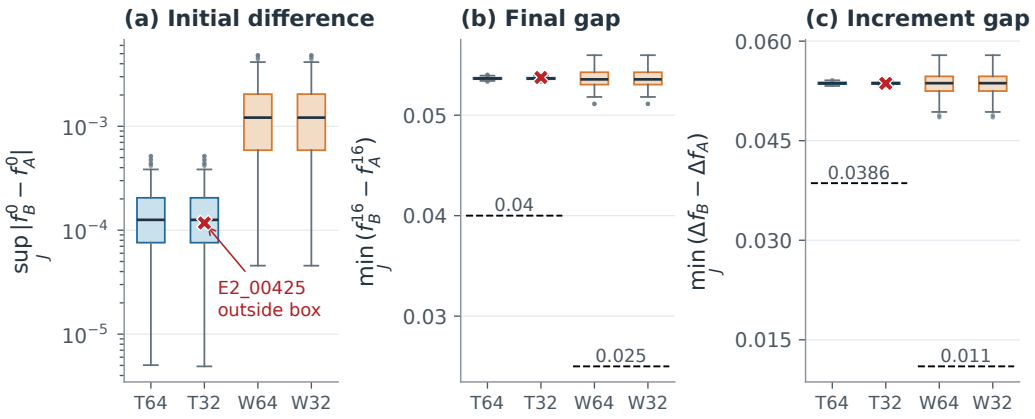


Figure 2: Independent perturbations: initial difference, final gap, and training-increment gap for all 1024 E2 pairs. Here $\Delta f = f^{16} - f^0$; T/W denote tight/wide boxes and 64/32 the precision. Boxes show medians and quartiles, with 1.5-IQR whiskers and all outliers. Dashed references are the final thresholds $(0.04, 0.025)$ and increment thresholds $(0.0386, 0.011)$. Red crosses retain the out-of-box record `E2_00425`, excluded only from the eligible-pair table.

E1: histories at matched empirical states. We use the prescribed-state construction of Theorem C.2. For GF, the widths are $m = 3, 4, 6, 8, 12$, with eight rational line witnesses per width and three target empirical states. Each witness consists of m lines. All 120 GF and 48 executed GD trajectories matched their required histories. The executed GD widths are 3 and 4; the 24 planned width-6 runs remain unrun because their step counts exceed the frozen cap. Appendix M.6 retains the width-12 timing warning and high-precision follow-up; neither provides a rigorous ODE enclosure.

E2: independent perturbations to the robust example. For the second experiment, we fix training rate $\eta = 0.005$ and $N = 16$, under different precision and parameter settings, each carrying out 64 perturbations of the pair θ_A and θ_B in Example 1, with 1024 executed pairs of networks in total. Of these, 1023 satisfy their actual binary initialization/data boxes and update budget predicted in the theoretical studies: $511/512$ in the tight box and $512/512$ in the wide box (Table 1). The retained tight-box float32 record `E2_00425` has initial error $1049/10485760 > 10^{-4}$ and is excluded from the theorem-eligible denominator. Every eligible pair had no A collision and exactly one anchored B collision $(2, 3)$ in steps 8–15; Figure 2 and Table 1 report the initial, final, and increment differences. The independently audited maximum total update error over all E2 runs is $5.424 \times 10^{-7} < 10^{-6}$, including arithmetic effects as well as injected noise.

Table 1: E2 descriptive extrema over eligible pairs only. Initial reports the largest initial discrepancy; final, increment, and final squared $L^2([0, 1])$ report the smallest observed values. Cross-precision cases are paired; no pooled independence-based confidence interval is asserted. Values are rounded measurements.

ρ	Eligible/tried	Initial $\sup_J$	Final $\min_J$	Increment $\min_J$	Final L^2 sq.
10^{-4}	511/512	5.17×10^{-4}	0.05336	0.05325	5.69×10^{-4}
10^{-3}	512/512	4.83×10^{-3}	0.05112	0.04852	5.46×10^{-4}

E3: limits of the observation window. We now change the learning rate and total number of steps for the perturbed pairs as above to understand their effects. In the fixed-duration rate scan, all 32 pairs at $\eta = 0.04$ left the original mask/sign domain and had negative final $\min_J(f_B - f_A)$. At 128 steps, even the zero-weight probe group had $32/32$ domain exits and identical full anchored words within each pair; zero weight adds no effective training sample. The complete scans and controls are in Appendix M.5. These finite constructed instances do not establish the universal perturbation theorem, long-time separation, or a causal effect of collisions.

E4: two-dimensional and two-hidden-layer examples. We test Examples 2 and 3 with all 13 and 23 parameters trained, respectively. Each model has one nominal pair and 64 perturbed pairs. Within a pair, the networks share the data but have independent initialization and, when enabled, update perturbations. All 130 pairs meet their actual error budgets. Along $y = 0$, A has no tracked collision and B has one $(2, 3)$ collision. Final prediction and training-increment differences are positive throughout the fixed test rectangles (Table 2). The two-hidden-layer runs first cross a training activation boundary, then undergo the spatial collision. Their small nodewise diagnostic differences do not imply small differences between the two boundary-crossing times. Figure 3 shows the nominal histories; Appendix R gives the settings and diagnostic plots.

Table 2: E4: extrema across 65 pairs per model. Initial is the largest $\sup_J |f_B^0 - f_A^0|$. Final and Increment are the smallest values of $\min_J(f_B^N - f_A^N)$ and $\min_J(\Delta f_B - \Delta f_A)$, where $\Delta f = f^N - f^0$. The fixed rectangles J are given in Appendix R. Values are rounded, not new theorem bounds.

Model	Pairs	Initial	Final	Increment
Planar	65	5.68419×10^{-4}	0.050445	0.050223
Two hidden layers	65	7.48747×10^{-6}	0.022599	0.022596

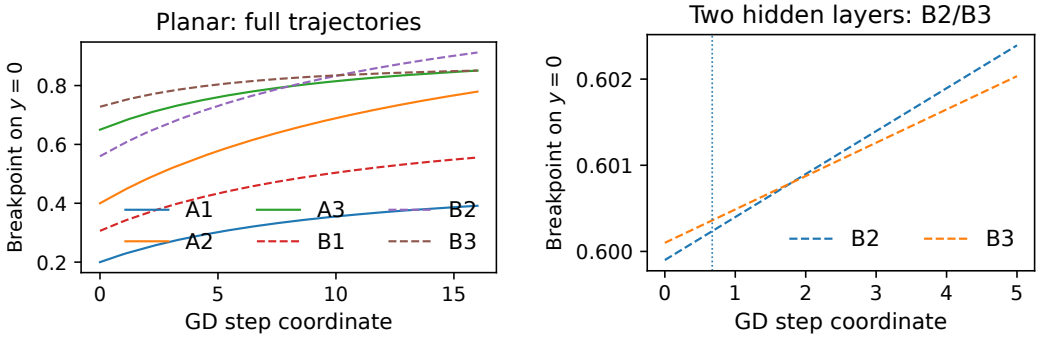


Figure 3: E4 nominal histories on $y = 0$. Left: all three tracked breakpoints in the planar pair (solid A, dashed B). Right: an early-time view of B2/B3 in the two-hidden-layer pair. The dotted line marks the lower unit's training activation change; the later intersection is a spatial breakpoint collision. Fractional steps refer to linear interpolation of parameter vectors.

E5: adding training samples. We use 8, 16 and 32 samples, with 16 matched data seeds per model and sample count, for 192 network pairs. The neurons mixed by Q are all three hidden units in the planar model and the first three upper units in the two-hidden-layer model. Added points either activate the whole block (*preserve*) or give it mixed signs (*split*) at initialization. Both groups use the same total training weight and optimization schedule. The 96 preserve pairs keep a common block pattern throughout training. The 96 split pairs already have different predictions at the added points before training; they are controls, not further identical-record examples. Preserve has a larger final rectangle gap in all 96 matched comparisons (Figure 4). Input locations also differ, so this comparison does not isolate a causal effect. Appendix R reports read-only evaluation without retraining and the retained sign-classification uncertainty near roots in 47 two-hidden-layer split pairs.

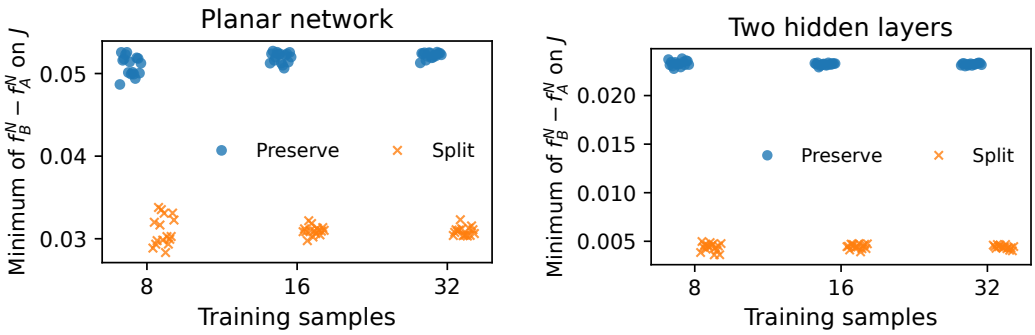


Figure 4: E5: final minimum prediction difference on each model's fixed rectangle. Each point is one network pair; horizontal offsets separate the 16 seeds in each group. Datasets are nested across sample counts and paired between groups. Split pairs are distinguishable at initialization.

7 Discussion

The main theorem and the examples show that the empirical training record need not determine the breakpoint history or the learned function away from the samples. E4 tests two-dimensional and two-hidden-layer examples. E5 shows that adding samples can preserve or break the common activation pattern, depending on their locations. These are selected constructions, not evidence of how often the effect occurs under standard random initialization. Its frequency on other data, its behavior over longer training windows, and its full range of breakpoint histories remain questions for further work.

REPRODUCIBILITY STATEMENT

The main text gives the argument for the main theorem; the appendices provide the detailed proofs and finite-certificate conditions. The original E1–E3 records, including unrun and out-of-scope cases, are described in Appendix M. E4–E5 have a separate archive of 322 pairs (Appendix R). A separate Lean development reports proofs for the earlier V10/V11 statements. Appendix L maps that reported scope to the present paper and distinguishes it from the unfinished formal hardness chain and the later extensions. The source package's `ARTIFACT_MAP.md` separates bundled evidence summaries from the larger code and audit archives. We do not claim end-to-end Lean verification of this manuscript.

AI USE STATEMENT

Generative AI tools assisted with research ideas and conceptual development, formulation and analysis of mathematical claims, proof drafting and critique, literature search, experimental design, implementation of checking programs, Lean development, and manuscript drafting and editing (especially the supplementary). Responsibility for the claims, citations, code, and any submitted empirical results remains with the authors.

REFERENCES

Igor Araujo, Alexander E. Black, Amanda Burcroff, Yibo Gao, Robert A. Krueger, and Alex McDonough. Realizable Standard Young Tableaux. arXiv:2302.09194v2, 2026. Version 2, 30 June 2026. First posted in 2023. https://arxiv.org/abs/2302.09194v2.

Yossi Arjevani and Michael Field. Analytic Characterization of the Hessian in Shallow ReLU Models: A Tale of Symmetry. In *NeurIPS 33*, 2020. Publication record.

Brian Wesley Bell, Michael Geyer, David Glickenstein, Amanda S. Fernandez, and Juston Moore. An Exact Kernel Equivalence for Finite Classification Models. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning; PMLR 221*, pp. 206–217, 2023. https://proceedings.mlr.press/v221/bell23a.html.

Joachim Bona-Pellissier, François Malgouyres, and François Bachoc. Geometry-induced Regularization in Deep ReLU Neural Networks. arXiv:2402.08269v2, 2026. Version 2, 5 January 2026. https://arxiv.org/abs/2402.08269v2.

Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004. https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 Theorem Prover and Programming Language. In *Automated Deduction – CADE 28, LNCS 12699*, pp. 625–635, 2021. https://link.springer.com/chapter/10.1007/978-3-030-79876-5_37.

J. Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden Symmetries of ReLU Networks. In *ICML; PMLR 202*, pp. 11734–11760, 2023. https://proceedings.mlr.press/v202/grigsby23a.html.

Moritz Grillo and Guido Montúfar. Most ReLU Networks Admit Identifiable Parameters. arXiv:2605.03601v2, 2026. Version 2, 19 May 2026. https://arxiv.org/abs/2605.03601v2.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, Springer Series in Computational Mathematics 8, second revised edition, 1993. Relevant passages checked in the corrected third printing (2008). https://doi.org/10.1007/978-3-540-78862-1.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature 585*, pp. 357–362, 2020. https://numpy.org/citing-numpy/.

Udo Hoffmann and Keno Merckx. A Universality Theorem for Allowable Sequences with Applications. arXiv:1801.05992v1, 2018. https://arxiv.org/abs/1801.05992v1.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *NeurIPS 31*, 2018. https://arxiv.org/abs/1806.07572.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics. In *ICLR*, 2021. https://arxiv.org/abs/2012.04728v2.

Colin Mallows and Robert J. Vanderbei. Which Young Tableaux Can Represent an Outer Sum? *Journal of Integer Sequences 18, Article 15.9.1*, 2015. https://cs.uwaterloo.ca/journals/JIS/VOL18/Vanderbei/vand3.pdf.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the Law and Follow the Flow: Conservation Laws for Gradient Flows. In *NeurIPS 36*, pp. 63210–63221, 2023. Publication record.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or Conservative? Conservation laws for ResNets and Transformers. arXiv:2506.06194v1, 2025. https://arxiv.org/abs/2506.06194v1.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, 2020. https://arxiv.org/abs/1910.09336.

Marco Nurisso, Pierrick Leroy, and Francesco Vaccarino. Topological obstruction to the training of shallow ReLU neural networks. In *NeurIPS 37*, pp. 35358–35387, 2024. Publication record.

Melissa E. O'Neill. PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation. Harvey Mudd College, technical report HMC-CS-2014-0905, 2014. https://www.cs.hmc.edu/tr/hmc-cs-2014-0905.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS 32*, 2019. Publication record.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica 19*, pp. 287–449, 2010. https://www.tuhh.de/ti3/paper/rump/Ru10.pdf.

Justin Sahs, Ryan Pyle, Aneel Damaraju, Josue Ortega Caro, Onur Tavaslioglu, Andy Lu, Fabio Anselmi, and Ankit B. Patel. Shallow Univariate ReLU Networks as Splines: Initialization, Loss Surface, Hessian, and Gradient Flow Dynamics. *Frontiers in Artificial Intelligence 5:889981*, 2022. Publication record.

The SciPy community. scipy.stats.bootstrap documentation. Official documentation, version 1.17.0, n.d. Accessed 22 September 2026. Method reference; the experiments use a custom implementation. Publication record.

Pierre Stock and Rémi Gribonval. An Embedding of ReLU Networks and an Analysis of Their Identifiability. *Constructive Approximation 57*, pp. 853–899, 2023. Theorem locators refer to the author version arXiv:2107.09370v5; journal volume published in 2023. https://link.springer.com/article/10.1007/s00365-022-09578-1.

Gerald Teschl. Ordinary Differential Equations and Dynamical Systems. American Mathematical Society, 2012. Graduate Studies in Mathematics, Vol. 140. author's errata.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17*, pp. 261–272, 2020. https://scipy.org/citing-scipy/.

Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient Dynamics of Shallow Univariate ReLU Networks. In *NeurIPS 32*, 2019. Publication record.

Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association 22(158)*, pp. 209–212, 1927. https://www.jstor.org/stable/2276774.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, Flat Minima, and the Conserved Quantities of Gradient Flow. In *ICLR*, 2023. https://arxiv.org/abs/2210.17216v2.

The appendices give the detailed arguments, certificate conditions and experimental protocols. The AI use statement applies to these materials. Formal and numerical verification cover only the scopes stated in Appendices L and R.

A ORIGINAL MODEL, EVENT DEFINITION AND EMPIRICAL STATE

For $m \geq 2$, train every coordinate of

$$f_\theta(x) = c + \sum_{j=1}^{m} a_j \,\mathrm{ReLU}(w_j x + b_j), \qquad L = \frac{1}{2}\{(f_\theta(0) - y_0)^2 + (f_\theta(1) - y_1)^2\}.$$

Write $u_j = w_j + b_j$, $q_j = w_j - b_j$, $t_j = -b_j/w_j$ and $r_i = f_\theta(i) - y_i$. The realization theorem uses the safe domain

$$w_j > 0, \quad b_j < 0, \quad u_j > 0, \quad a_j > 0, \quad r_0 < 0 < r_1. \tag{5}$$

Every breakpoint is in $(0, 1)$. Collisions there do not change either training sample's activation; the original loss is polynomial on this domain. The reduction below is for this fixed dataset; conservation for a given dataset and for arbitrary datasets are distinct notions (Marcotte et al., 2023, Definitions 2.2–2.4). Sample contacts and exits from the specified sign domain are stopping walls.

A complete reversal word begins with anchored left-to-right labels $1, \ldots, m$, consists of $\binom{m}{2}$ adjacent swaps, uses each unordered label pair exactly once, and ends in reverse order. A *stretchable* word is represented by the top-to-bottom ordering of affine lines over $[0, 1]$, with distinct simple interior pair intersections and strictly ordered endpoints. The full order of swaps is retained, not a commutation class (Hoffmann and Merckx, 2018). For GD we use its parameter-linear polygon and the half-open convention $(n, n + 1]$; an intermediate node event is assigned once but that node itself has no strict permutation.

The original gradients are

$$\dot{a}_j = -r_1 u_j, \quad \dot{w}_j = \dot{b}_j = -r_1 a_j, \quad \dot{c} = -(r_0 + r_1). \tag{6}$$

Set

$$D = \sum_j a_j u_j, \quad E = \sum_j (u_j^2 + 2a_j^2), \quad \chi_j = \frac{2a_j^2 - u_j^2}{q_j^2}.$$

Direct differentiation with respect to original Euclidean parameters gives

$$f_{\mathrm{train}} = (c, c + D), \qquad K_{\mathrm{emp}} = \begin{pmatrix} 1 & 1 \\ 1 & 1 + E \end{pmatrix}. \tag{7}$$

The closed GF equations are

$$\dot{c} = -(2c + D - y_0 - y_1), \qquad \dot{D} = -r_1 E, \qquad \dot{E} = -8 r_1 D. \tag{8}$$

For an original GD step of size η, $\gamma = \eta r_1$ and

$$\begin{aligned} u_j^+ &= u_j - 2\gamma a_j, & a_j^+ &= a_j - \gamma u_j, & q_j^+ &= q_j, \\ c^+ &= c - \eta(2c + D - y_0 - y_1), \\ D^+ &= (1 + 2\gamma^2) D - \gamma E, & E^+ &= (1 + 2\gamma^2) E - 8\gamma D. \end{aligned} \tag{9}$$

The prediction and kernel dynamics close in (c, D, E) while the paths are safe. Equality of the full record (2) also requires matching $S_q = \sum_j q_j^2$, conserved by both GF and GD. Thus equal initial (c, D, E, S_q) gives equal full records on safe GF paths, or on safe GD polygons with common rates; the spectral equality follows from Appendix B. On a polygon segment replace η, γ by their fractional values. The norm identities are

$$\|\nabla L\|^2 = r^\top K_{\mathrm{emp}} r, \qquad \|\theta\|^2 = c^2 + \tfrac{1}{2}\big(E + \sum_j q_j^2\big).$$

The residual system is $\dot{r}_0 = -r_0 - r_1$, $\dot{r}_1 = -r_0 - (1 + E) r_1$. Variation of constants for $(-r_0, r_1)$ preserves strict positivity on each finite safe interval, so $d\tau = r_1 \, ds$ is a strictly increasing clock.

B EXACT ORTHOGONAL FIBERS AND THEIR CONSEQUENCES

The empirical fibers admit an explicit orthogonal description. Their constraints differ from the neuronwise conserved levels studied by Nurisso et al. (2024, Sections 4–5 and Appendix E), whose bias extension is included in that appendix. Linear loss symmetries provide the relevant perspective (Zhao et al., 2023, Section 3.3); the data-dependent nonlinear action of their Theorem 4.1 is not used here. Our direct calculation establishes the isometries and spectral equalities for the present empirical-moment fibers.

Lemma B.1 (Orthogonal mode coordinates)**.** *Define, neuronwise,*

$$z_{+,j} = \frac{a_j}{\sqrt{2}} + \frac{w_j + b_j}{2}, \quad z_{-,j} = \frac{a_j}{\sqrt{2}} - \frac{w_j + b_j}{2}, \quad \bar{q}_j = \frac{w_j - b_j}{\sqrt{2}}.$$

This is an orthogonal change of the original (a_j, w_j, b_j) coordinates. With $z_+, z_- \in \mathbb{R}^m$,

$$\begin{aligned} D &= (\|z_+\|^2 - \|z_-\|^2)/\sqrt{2}, & E &= 2(\|z_+\|^2 + \|z_-\|^2), & & (10)\\ \dot{z}_+ &= -\sqrt{2} r_1 z_+, & \dot{z}_- &= \sqrt{2} r_1 z_-, & \dot{\bar{q}} &= 0,\\ z_+^+ &= (1 - \sqrt{2}\gamma) z_+, & z_-^+ &= (1 + \sqrt{2}\gamma) z_-, & \bar{q}^+ &= \bar{q}. \end{aligned}$$

Proof. The three rows $(1/\sqrt{2}, 1/2, 1/2)$, $(1/\sqrt{2}, -1/2, -1/2)$ and $(0, 1/\sqrt{2}, -1/\sqrt{2})$ are orthonormal. Expanding the two squares proves the moment identities; applying this matrix to (6) and (9) gives the rest. □

Proposition B.2 (Regular fiber and fixed path isometry)**.** *Fix c, the vector q, and moments D, E with $E > 2\sqrt{2}|D|$. The algebraic fixed-moment fiber is*

$$S^{m-1}\left(\sqrt{\frac{E + 2\sqrt{2}D}{4}}\right) \times S^{m-1}\left(\sqrt{\frac{E - 2\sqrt{2}D}{4}}\right), \tag{11}$$

with dimension $2m - 2$. Any two points of it are related by a fixed orthogonal transformation T acting as independent rotations/reflections of z_+, z_- and as identity on q, c. For safe GF paths from these points, $\widetilde{\theta}(s) = T\theta(s)$ at common physical times. For GD with common rates, the same relation holds at every common node and interpolation fraction for which both paths are safe.

Proof. The squared radii follow by solving (10). Both are positive. Each orthogonal group is transitive on its sphere, so choose two fixed orthogonal matrices mapping the initial mode vectors. Empirical predictions and the polynomial loss depend on the modes only through their two squared norms and c. Thus T preserves this polynomial and the original Euclidean metric. Equivalently, the explicit mode updates in Lemma B.1 commute with T. Equal empirical states provide the same scalar residual, so uniqueness or induction proves the path relation. Linear interpolation commutes with T. □

Domain restriction. The polynomial extension is invariant on its whole algebraic domain, but the original ReLU loss equals it only while the sample masks remain valid. We do not assert that every point in the orbit, or an arbitrary rotation of a safe point, is safe. The statement compares two separately certified safe trajectories. It is not a global input-function symmetry of ReLU networks.

Corollary B.3 (Additional observations that still do not distinguish histories)**.** *On corresponding safe paths in Proposition B.2, full loss Hessians are orthogonally conjugate:*

$$\nabla^2 L(\widetilde{\theta}(s)) = T \nabla^2 L(\theta(s)) T^\top.$$

Their spectra, all cross-time parameter inner products, and all cross-time empirical gradient kernels $\langle \nabla f_\theta(i, s), \nabla f_\theta(k, t) \rangle$, $i, k \in \{0, 1\}$, agree. Hessian matrices or eigenvectors in the original fixed coordinate basis need not agree.

Proof. Differentiate the fixed polynomial identities $L(T\theta) = L(\theta)$ and $f(T\theta, i) = f(\theta, i)$ near each pair of safe points. Orthogonality gives gradient and Hessian conjugacy. The same *fixed* T applies at every time, proving the cross-time identities. This argument would be invalid for an arbitrary nonorthogonal loss-preserving reparameterization. □

For example, scaling a by 2 and u by $1/2$ preserves the prediction product au but generally changes $u^2 + 2a^2$. Equality of one loss value does not give the empirical quotient, still less a Euclidean optimizer equivalence.

An explicit checkable Hessian formula is useful for the audit. Put $r = r_1$. In the original $3m + 1$ parameters,

$$\det(\lambda I - \nabla^2 L) = \lambda^m(\lambda^2 - 2r^2)^{m-1}\left[\lambda^3 - (E+2)\lambda^2 + (E - 4Dr - 2r^2)\lambda + 4Dr + 4r^2\right]. \tag{12}$$

Indeed q contributes m zero modes. Tangents to the two spheres contribute $\sqrt{2}r$ and $-\sqrt{2}r$, each $m - 1$ times. Writing $A = \|z_+\|^2$, $B = \|z_-\|^2$, the residual radial block is

$$\begin{pmatrix} 2A + \sqrt{2}r & -2\sqrt{AB} & \sqrt{2A} \\ -2\sqrt{AB} & 2B - \sqrt{2}r & -\sqrt{2B} \\ \sqrt{2A} & -\sqrt{2B} & 2 \end{pmatrix}.$$

Its characteristic polynomial is the cubic in (12). No convexity or distinct-eigenvalue assumption is used. The spectral equality follows from the fixed isometry; see Zhao et al. (2023); Arjevani and Field (2020) for broader symmetry-based analyses.

C An open set of empirical states with the full history language

We now combine the line representation with the empirical-state equations. There are three tasks: fit the prescribed moments, preserve the chosen crossing order, and prove that the original optimizer reaches every crossing while the masks remain valid. The common-offset construction handles the first two; the timing bounds below handle the third. Throughout, the target is chosen *before* the word. The common trace is shared among words at one target, not across different targets.

Fix $m \geq 2$ and positive rational tolerances r, δ, T_0. Set $r_* = \min(r, 1/4)$ and choose rational

$$0 < \varepsilon \leq \min\{10^{-3}, r_*/200, \delta/12, mT_0/4\}. \tag{13}$$

Let $h_0 = 1/7$, $v_0 = 5/7 + \varepsilon$, $p_0 = h_0 v_0$ and $s_0 = h_0^2 + 2v_0^2$. Define

$$\mathcal{B}_{m,\varepsilon} = \left\{(c_*, D_*, E_*) : |c_* + 1| < \varepsilon/100,\ \left|\frac{D_*}{49m} - p_0\right| < \varepsilon/100,\ \left|\frac{E_*}{49m} - s_0\right| < \varepsilon/100\right\}. \tag{14}$$

The bounds are normalized empirical moments, not equal ambient-coordinate radii. The target is fixed before selecting a history.

C.1 Moment correction with explicit uniform margins

Take a rational line witness of a stretchable complete reversal on $[0, 1]$. Rationality can be arranged because, after eliminating pair times, the strict simple-word conditions are open in the line coefficients. A common positive scale and common affine subtraction give centered coefficients x_j, y_j with

$$\overline{x} = \overline{y} = 0, \qquad |x_j|, |y_j| \leq 1/2.$$

They preserve every pair intersection and its clock. Define the means $\mu_x = \overline{x^2}$, $\mu_y = \overline{y^2}$, $\mu_{xy} = \overline{xy}$; the first two are at most $1/4$ and $|\mu_{xy}| \leq 1/4$.

For a prescribed target in (14), put

$$\begin{aligned} &p = D_*/(49m) + \varepsilon^3\mu_{xy}, && s = E_*/(49m) - \varepsilon^4\mu_x - 2\varepsilon^2\mu_y, && \\ &R = \sqrt{s^2 - 8p^2}, && V = \sqrt{(s+R)/4}, && H = p/V. \end{aligned} \tag{15}$$

Use these in the original initialization

$$\begin{aligned} &h_j = H + \varepsilon^2 x_j, \qquad v_j = V - \varepsilon y_j, \qquad q_j = 7, \\ &a_j = 7v_j, \quad w_j = \tfrac{7}{2}(1 + h_j), \quad b_j = \tfrac{7}{2}(h_j - 1), \quad c = c_*. \end{aligned} \tag{16}$$

Lemma C.1 (Open-target lift)**.** *The radicals in* (15) *are positive and select a single smooth branch throughout the stated box. The lift satisfies exactly* $(c, D, E) = (c_*, D_*, E_*)$*. Moreover,*

$$|V - v_0| < \varepsilon/100, \quad |H - h_0| < \varepsilon/50, \quad \|\theta^0 - \bar{\theta}_m\|_\infty < 13\varepsilon, \quad \tfrac{6}{5}\varepsilon < \chi_j - 1 < 5\varepsilon, \tag{17}$$

where $\bar{\theta}_m$ *has every* $(a_j, w_j, b_j) = (5, 4, -3)$ *and* $c = -1$*. In the initial line representation,*

$$h_j - 2v_j(\varepsilon z/2) = H - \varepsilon V z + \varepsilon^2(x_j + y_j z). \tag{18}$$

Thus the full line word is unchanged.

Proof. We first control the root branch, then check the two moment identities and the safety margins. The branch estimates must be uniform in the witness: otherwise different words could require different empirical target boxes.

Let $R_0 = 2v_0^2 - h_0^2 \geq 1$. We have $s_0 < 1.05$, $p_0 < 0.103$, and

$$|p - p_0| < 0.011\varepsilon, \qquad |s - s_0| < 0.011\varepsilon.$$

Expanding the two discriminants gives

$$|(s^2 - 8p^2) - R_0^2| \leq 2s_0|s - s_0| + |s - s_0|^2 + 16p_0|p - p_0| + 8|p - p_0|^2 < \varepsilon/20.$$

Consequently $R > 0.99$, $V > 0.7$, and $|R - R_0| < \varepsilon/39$. Therefore

$$|V^2 - v_0^2| \leq \tfrac{1}{4}(|s - s_0| + |R - R_0|) < \varepsilon/100, \qquad |V - v_0| < \varepsilon/100.$$

Using $H = p/V$ then yields

$$|H - h_0| \leq (|p - p_0| + h_0|V - v_0|)/V < \varepsilon/50.$$

We can now verify the moments without further estimates. The branch satisfies $HV = p$ and $H^2 + 2V^2 = s$, by $2(V^2)^2 - sV^2 + p^2 = 0$. Centering eliminates the terms linear in x_j or y_j, leaving

$$\frac{D}{49m} = HV - \varepsilon^3\mu_{xy} = \frac{D_*}{49m}, \qquad \frac{E}{49m} = H^2 + 2V^2 + \varepsilon^4\mu_x + 2\varepsilon^2\mu_y = \frac{E_*}{49m}.$$

Equation (18) is direct expansion.

It remains to locate the lift relative to the balanced center. Write $v_j = 5/7 + \alpha_j$, $h_j = 1/7 + \beta_j$. The preceding bounds imply $0.49\varepsilon < \alpha_j < 1.51\varepsilon$ and $|\beta_j| < \varepsilon/40$. Hence

$$\chi_j - 1 = \frac{20}{7}\alpha_j + 2\alpha_j^2 - \frac{2}{7}\beta_j - \beta_j^2.$$

Bounding the negative terms gives a lower bound greater than 1.39ε, hence greater than $6\varepsilon/5$. Bounding positive terms gives an upper bound below $4.33\varepsilon < 5\varepsilon$. The initial a displacement is below 10.57ε, those of w, b below $7\varepsilon/80$, and that of c below $\varepsilon/100$. These imply the stated 13ε bound. All inequalities have strict slack. □

For rational target and witness, this lift has real algebraic coordinates of degree at most four. Rational target moments do not imply a rational exact lift. No polynomial encoding bound in *word length* is asserted.

C.2 Actual optimizers and common time

Theorem C.2 ($\mathrm{S}_{\mathrm{open}}$: complete histories on every target fiber)**.** *For every target* $(c_*, D_*, E_*) \in \mathcal{B}_{m,\varepsilon}$*, every stretchable complete reversal has a realization from* (16) *by the original GF, with the following properties. The entire path stays within radius* r *of* $\bar{\theta}_m$*, satisfies* (5)*, and has* $0 < \chi_j - 1 < \delta$*. All words are completed on the same physical interval* $[0, S_*]$*, with* $0 < S_* < T_0$*. On that interval, every realization has the same empirical prediction, loss, empirical NTK, gradient norm, total parameter norm, and full loss-Hessian spectrum.*

For any finite collection of words and any rational $\eta_{\max} > 0$*, one constant rational learning rate and one integer step count work for all of them under original GD, with the same whole-polygon equalities and safety. Each half-open step contains at most one collision, initial and final nodes are off collision walls, and total physical duration is below* T_0*. Initializations are real; for rational targets they may be chosen real algebraic.*

GF proof. The line witness already fixes the desired event order. To turn it into an optimizer trajectory, we first obtain a safe interval in physical time and then show that the residual clock traverses the entire witness interval before that time ends.

In the closed radius-$1/4$ box about $\bar{\theta}_m$, original quantities obey

$$a_j \in [19/4, 21/4], \quad u_j \in [1/2, 3/2], \quad -r_0 \in [3/4, 5/4], \quad m \le r_1 \le 10m, \quad \|\nabla L\|_\infty \le 60m.$$

The lower residual bound is $r_1 \ge (19m - 10)/8 \ge m$. Suppose the path were to exit the box before physical time ε/m. Up to that first exit, the gradient bound and initial distance $< 13\varepsilon$ would keep its distance below $73\varepsilon < r_*$, a contradiction. The polynomial vector field is smooth near this compact safe box. Local existence, uniqueness, and compact-set continuation therefore extend the original solution through that time (Teschl, 2012, Theorem 2.2, p. 38; Corollary 2.15, p. 52). We have therefore established safety before using the residual-time solution.

In residual time, $u' = -2a, a' = -u, q' = 0$, and

$$\xi = \tanh(\sqrt{2}\tau)/\sqrt{2}, \qquad \frac{u_j(\tau)}{q_j} = \frac{h_j - 2v_j\xi}{\sqrt{1 - 2\xi^2}}.$$

Reaching $\xi = \varepsilon/2$ requires $\tau_\varepsilon = \operatorname{artanh}(\varepsilon/\sqrt{2})/\sqrt{2} < \varepsilon$. Since $\dot{\tau} = r_1 \ge m$, it is reached before physical time ε/m. Every target clock lies strictly inside this interval, and (18) preserves its crossing order. Nonzero line slope difference and positive clock speed prove transversality; each pair is listed once, excluding extra pair events. Conservation of χ_j proves its bounds.

Finally, the target moments give the same aggregate solution (8) for every word. Hence all paths have the same r_1, the same residual clock as a function of physical time, and the same hitting time S_*. This is stronger than reaching the same formal line clock at unrelated times. The remaining observation equalities follow from (7) and Corollary B.3. The kernel is evolving throughout: $\dot{E} = -8r_1 D < 0$. □

Direct GD proof. The discrete argument has the same safety requirement, but its event order comes from the exact block product rather than approximation of GF. We need an upper bound on each clock step to separate events, and a lower bound to ensure completion.

Choose one rational geometric witness for each word in the finite collection. Let $g_* > 0$ be the minimum of all its normalized event gaps, including distances to $0, 1$. Set

$$\eta = \min\{\varepsilon g_*/(40m), \eta_{\max}/2\}, \qquad K = \lceil \varepsilon/(m\eta) \rceil.$$

Since $g_* \le 1$, $K\eta < 2\varepsilon/m$. Discrete first-exit induction and the same safe-box gradient bound give distance below $13\varepsilon + 60mK\eta < 133\varepsilon < r_*$ for every node and every interpolated segment. Hence $m \le r_{1,n} \le 10m$,

$$0 < \gamma_n \le \varepsilon g_*/4 \le \varepsilon/4, \qquad \sum_{n<K} \gamma_n < 20\varepsilon.$$

Actual original updates give $\chi_{j,n+1} = (1 - 2\gamma_n^2)\chi_{j,n}$ and

$$\prod_{n<K} (1 - 2\gamma_n^2) \ge 1 - 2\sum_{n<K} \gamma_n^2 \ge 1 - 10\varepsilon^2.$$

The same lower bound applies to every partial product and partial segment. Therefore

$$\chi_j > (1 + 6\varepsilon/5)(1 - 10\varepsilon^2) > 1 + \varepsilon, \qquad \chi_j < 1 + 5\varepsilon < 1 + \delta.$$

Multiplication of actual hidden updates yields common $P_0 = 1, Q_0 = 0$ with

$$P_{n+1} = P_n + 2\gamma_n Q_n, \quad Q_{n+1} = Q_n + \gamma_n P_n, \quad \xi_n = Q_n/P_n,$$

$$\xi_{n+1} - \xi_n = \frac{\gamma_n(1 - 2\xi_n^2)}{1 + 2\gamma_n\xi_n} > 0.$$

The same clock is strictly increasing within each parameter segment. First, its increment is at most $\varepsilon g_*/4$, while consecutive target clocks are separated by at least $\varepsilon g_*/2$. This rules out two events in a half-open step. Second, until ξ_n reaches $\varepsilon/2$, each increment is at least $m\eta/2$. If the endpoint had

not been reached after K steps, summing this bound would give $\xi_K \geq \varepsilon/2$, a contradiction. Thus every pair crosses. The initial and final clocks lie strictly outside the event set; an intermediate node collision belongs to the preceding half-open step.

Equal target moments and common η give the same aggregate sequence (9), including its fractional-step version. Fixed isometries yield the Hessian spectra and cross-time assertions. No GF shadowing is needed. Finally $K\eta < 2\varepsilon/m \leq T_0/2 < T_0$ and $\eta < \eta_{\max}$. □

C.3 Relative robustness and precise quantifiers

Corollary C.3 (Nonempty relative-open history cells)**.** *For each target in* (14) *and each stretchable complete reversal, the safe fixed-q, c, D, E fiber contains a nonempty relatively open set of GF initializations realizing that word on the common interval $[0, S_*]$. The fiber is regular of dimension $2m - 2$.*

Proof. The target has $E_* > 2\sqrt{2}|D_*|$ by the same discriminant bound, so Proposition B.2 applies. One can also see regularity in original hidden coordinates from

$$\det \begin{pmatrix} u_j & a_j \\ 4a_j & 2u_j \end{pmatrix} = -2q_j^2\chi_j \neq 0.$$

At the constructed point, initial order, pair slopes, all distinct interior event clocks, endpoint clearance, and compact safety/defect bounds are strict. Small perturbations within the regular fiber preserve these strict finite comparisons and margins. The original flow has continuous dependence on its initial state (Teschl, 2012, Theorem 2.9, p. 44). Its aggregate path and S_* remain identical on the fiber. Thus a relative neighborhood has exactly the same full word. □

The empirical box is open in three aggregate coordinates. An exact fiber is lower-dimensional in parameter space. Relative openness does not give a uniform lower bound on cell volume, a nonvanishing random-initialization probability, or a fixed-precision hardness result. The box shrinks as $\varepsilon \to 0$, and history margins may also shrink. At a fixed width the word set is finite, so one witness per realizable word supplies a common GD schedule for the whole language; this finite minimum need not have an efficiently controlled size. We make no unconditional relative-open *node-event* GD claim.

C.4 The open parameter set used in Theorem 3.1

The preceding theorem fixes $q_j = 7$. To justify the open set of initial *parameters* in Theorem 3.1, the comparison must also match the initial parameter norm. We give that step explicitly; we do not treat the fixed-q fiber as an ambient open set.

Fix a rational $0 < \epsilon \leq 10^{-4}$ for this subsection and set

$$h_0 = 1/7, \qquad v_0 = 5/7 + \epsilon, \qquad p_0 = h_0 v_0, \qquad s_0 = h_0^2 + 2v_0^2.$$

We reuse the normalized scalar estimates of Lemma C.1 with this ϵ; safety of the open parameter set is checked below. Write

$$q_*(\theta) = \left(\frac{1}{m} \sum_j (w_j - b_j)^2 \right)^{1/2}.$$

Let Y consist of the points with $\|\theta - \bar{\theta}_m\|_\infty < 1/8$ satisfying

$$|q_* - 7| < \epsilon, \quad |c + 1| < \epsilon/100, \quad \left| \frac{D}{mq_*^2} - p_0 \right| < \epsilon/100, \quad \left| \frac{E}{mq_*^2} - s_0 \right| < \epsilon/100.$$

This is open, and it contains the point with every $(a_j, w_j, b_j) = (5 + 7\epsilon, 4, -3)$ and $c = -1$. The radius-$1/8$ condition also places Y inside the strict training region.

For a given $\theta \in Y$ and a centered line witness, apply Lemma C.1 with normalized targets $D/(mq_*^2)$ and $E/(mq_*^2)$. Replace the factor 7 in (16) by q_*. The same scalar root bounds apply. Centering gives

$$\frac{D(\theta_\omega)}{mq_*^2} = HV - \epsilon^3\mu_{xy} = \frac{D(\theta)}{mq_*^2}, \qquad \frac{E(\theta_\omega)}{mq_*^2} = H^2 + 2V^2 + \epsilon^4\mu_x + 2\epsilon^2\mu_y = \frac{E(\theta)}{mq_*^2}.$$

Also $c(\theta_\omega) = c(\theta)$ and $\sum_j q_j(\theta_\omega)^2 = mq_*^2 = \sum_j q_j(\theta)^2$. Scaling the radius-$13\epsilon$ lift of the lemma by $q_*/7$ gives

$$\|\theta_\omega - \bar{\theta}_m\|_\infty < 13\epsilon(1 + \epsilon/7) + 5\epsilon/7 < 14\epsilon < 1/8.$$

Hence $\theta_\omega \in Y$. The two mode vectors z_+, z_- and the vector $\bar{q}$ in Lemma B.1 have matching norms. Separate fixed orthogonal maps on these three vectors give an isometry between the initializations. The fixed-mask polynomial loss does not depend on $\bar{q}$, so this isometry preserves the original flow, GD updates, and Hessian spectra while both paths remain safe. The norm identity in Appendix A supplies the remaining record component.

Safety is uniform in this construction. In the radius-$1/4$ box, $m \le r_1 \le 10m$ and $\|\nabla L\|_\infty \le 60m$, as in the preceding proof. For GF, an exit before ϵ/m would require a displacement at least $1/8$, whereas the gradient bound allows at most $60\epsilon < 1/8$. Thus both paths stay safe until the common residual clock reaches $\xi = \epsilon/2$. The line comparisons then give the required word with the common finite endpoint.

For GD, choose η and K from the preceding direct proof for one witness or any finite family of witnesses. The total duration is $K\eta < 2\epsilon/m$, and $1/8 + 120\epsilon < 1/4$ keeps all nodes and segments in the safe box. The exact recurrences for D, E and the common discrete clock are independent of the individual q_j. The event-spacing and clock-completion bounds therefore apply unchanged. This establishes exact record equality for the chosen common schedule, not for arbitrary learning rates or an unlimited observation window.

D What complete function observations do recover

The spline representation exposes the moving knots and slope jumps (Sahs et al., 2022, Section 2.1). The empirical theorem is not a theorem of intrinsic unobservability of the history. The following contrasting statement also rules out a merely arbitrary relabeling explanation for *every pair* of different GF words, not just checked examples.

Proposition D.1 (Full function-trace recovery)**.** *Consider a finite-width GF path on $[0, S]$ with $a_j, w_j > 0$, all breakpoints in $(0, 1)$, and only finitely many simple transverse pair collisions strictly inside the interval. No endpoint or simultaneous collision occurs. The exact function trace $(s, x) \mapsto f_{\theta(s)}(x)$ for all $s \in [0, S]$, $x \in (0, 1)$, together with the initial spatial anchoring of neuron labels, uniquely determines the complete labeled breakpoint history.*

Proof. The function trace first reveals the moving knot locations. The only ambiguity is how to continue the labels when two locations meet; transversality and smoothness of the original flow resolve it.

For each fixed time, distributional differentiation in the input gives

$$\partial_x^2 f_{\theta(s)} = \sum_{j=1}^{m} a_j(s) w_j(s)\, \delta_{t_j(s)}.$$

All coefficients are strictly positive. Away from collision times the support therefore determines every breakpoint without cancellation. The initial ordering assigns the labels. Before the first collision, continuity assigns each label uniquely to its isolated moving support point.

At a simple pair collision only two support branches meet. Transversality says their time derivatives are different. The actual parameter flow is smooth at this event, because neither sample activation changes; hence each $t_j(s)$ is continuously differentiable. From the exact support curves on the two sides, each incoming derivative can be matched to exactly one outgoing derivative. The incorrect alternative would splice different slopes and fail to be C^1. This determines the continuation of both labels. Other supports remain separated. Repeating across the finite set of collisions reconstructs the whole word. Positivity also ensures the count of distinct support points drops at each collision rather than losing a kink by cancellation. □

Consequently any two different GF words constructed in Theorem C.2 have different full input-function traces somewhere, despite all the empirical and spectral equalities. This does not identify the full weights, quantify a prediction gap, or supply a stable finite-query reconstruction algorithm.

The statement is about GF. We do not extend its C^1 branch-matching argument to arbitrary GD node collisions, where parameter velocities can jump. Positivity and transverse/simple events are explicit sufficient assumptions, not a claimed characterization of every recoverable case.

E EXACT FINITE-CLOCK REACHABILITY AND A FORWARD ALGORITHM

The line representation identifies candidate events; this section determines which ones the original flow actually reaches. The criterion also allows fixed negative output signs, subject to positive slopes, interior knots, and the strict residual cone. Here a_j, u_j, q_j on right-hand sides denote initial values.

Theorem E.1 (Finite-clock criterion)**.** *Let $q_j = w_j - b_j > 0$, $0 < u_j < q_j$, $a_j \neq 0$, and $r_0 < 0 < r_1$ initially. Put*

$$D = \sum_j a_j u_j, \quad E = \sum_j (u_j^2 + 2a_j^2), \quad \Delta = y_1 - y_0,$$

$$d = 1 - 2\xi^2, \quad U_j = u_j - 2a_j\xi, \quad A_j = a_j - u_j\xi, \quad G = (D - \Delta) - E\xi + 2(D + \Delta)\xi^2.$$

For $0 < \xi_* < 1/\sqrt{2}$, the corresponding clock is reached by actual GF at a finite physical time strictly before every sample or output-sign wall if and only if

$$G(\xi_*) > 0, \qquad U_j(\xi_*) > 0, \qquad q_j^2 d(\xi_*) - U_j(\xi_*)^2 > 0, \qquad a_j A_j(\xi_*) > 0 \quad \forall j.$$

Initial knot coincidences are allowed for this reachability claim; a strict event word requires separate genericity.

Proof. Necessity follows from the explicit hidden solution. Sufficiency requires more: endpoint inequalities must control the entire hidden path, and the residual clock must keep advancing in physical time. We establish these two facts separately.

On a safe sample stratum the original gradients imply $u' = -2a$, $a' = -u$, $q' = 0$ in residual time. The residual cone is strictly invariant on finite safe intervals by the cooperative equations in $(-r_0, r_1)$ and variation of constants. Solving the hidden linear system gives

$$u_j(\tau) = \frac{U_j(\xi)}{\sqrt{d}}, \quad a_j(\tau) = \frac{A_j(\xi)}{\sqrt{d}}, \quad \xi = \tanh(\sqrt{2}\tau)/\sqrt{2}.$$

The prediction difference and residual gap are

$$D(\xi) = \frac{D(1 + 2\xi^2) - E\xi}{d}, \qquad g(\xi) := r_1 - r_0 = D(\xi) - \Delta = G(\xi)/d.$$

These identities immediately prove necessity, including strictness at the endpoint.

For sufficiency, U_j and $a_j A_j$ are affine and positive at both endpoints $0, \xi_*$. The sample margin satisfies

$$\frac{d^2}{d\xi^2}(q_j^2 d - U_j^2) = -4q_j^2 - 8a_j^2 < 0.$$

A concave function positive at both endpoints is positive throughout. Thus the entire formal hidden path, not just its endpoint, has $0 < u_j(\tau) < q_j$ and the prescribed nonzero output sign. The denominator also stays positive. Along it,

$$\frac{dD(\tau)}{d\tau} = -\sum_j (u_j(\tau)^2 + 2a_j(\tau)^2) \leq 0.$$

The inequality is strict when the neuron family is nonempty; nonincrease suffices also for the empty family. Therefore g has a positive lower bound g_* on the compact residual interval $[0, T]$ corresponding to ξ_*, and the hidden variables and $K = 1 + E(\tau)$ have finite bounds, say $K \leq K_*$. This establishes uniform hidden safety independently of any assertion of physical completion.

We now return to physical time. A positive lower bound on the gap g prevents the positive residual from decaying to zero before the target clock. While the original solution has reached residual time at most T,

$$\dot{r}_1 = g - (K + 1)r_1, \qquad r_1(s) \geq \rho_* := \min\{r_1(0), g_*/(K_* + 1)\} > 0.$$

The last bound follows by scalar comparison at the constant barrier. Also the strict cone bounds the bias by $y_1 - D(\tau) < c < y_0$. Thus the whole original parameter state is bounded. On a putative finite maximal interval before reaching T, the hidden variables have uniform distance from their walls and r_1 has the stated lower bound. Moreover $-r_0(s) \geq (-r_0(0))e^{-s} > 0$ on every finite such interval. The polynomial vector field has no finite breakdown, and compact-set continuation extends the solution (Teschl, 2012, Corollary 2.15, p. 52). Since $\dot{\tau} = r_1 \geq \rho_*$, it reaches T by time T/ρ_*. This proves actual physical reachability, not only solvability of the reparameterized equations. □

If the first positive zero ξ_∞ of G precedes all hidden walls and $1/\sqrt{2}$, each smaller clock is attained, but the boundary itself cannot be attained at finite physical time because it would have $r_1 - r_0 = 0$ in the strict cone. The hidden variables remain bounded on the closed clock interval, so physical continuation holds indefinitely. The clock converges to ξ_∞; a smaller interior limit would contradict the positive-barrier argument. Hence $0 < r_1 < g \to 0$ and both residuals vanish. A candidate pair collision exactly at that limit is *not* a finite event. A pair collision on a terminal parameter wall is likewise excluded from the regular interior language.

Corollary E.2 (Rational forward prediction)**.** *For supplied rational initial parameters and targets satisfying the strict domain, the maximal generic finite safe GF history is computable in polynomial bit complexity. A direct algorithm uses $O(m^3 + m^2 \log m)$ rational operations.*

Proof. Compute $h_j = u_j/q_j$ and $v_j = a_j/q_j$. For each nonparallel pair compute the rational candidate $\xi_{ij} = (h_i - h_j)/(2(v_i - v_j))$. Sort the positive candidates and retain precisely those with $d > 0$ passing Theorem E.1. Equal retained clocks or repeated initial knots are reported as non-generic rather than arbitrarily ordered. Each retained simple line crossing exchanges adjacent lines. The forward word is recovered by these swaps. There are $O(m^2)$ candidates, each with $O(m)$ fixed-degree rational tests. All intermediate expressions have polynomial bit length in the supplied inputs. No root isolation or ODE tolerance is needed. The algorithm does not compute physical event times or the word up to a prescribed physical horizon. □

F THE EXACT DISCRETE CLOCK

Kunin et al. (2021, Appendix I) derive discrete changes of symmetry-induced conserved quantities. Marcotte et al. (2025, Proposition 5.1) bound the expected absolute drift of a conserved function h. Their assumptions bound $\nabla^2 h$ and the second moment of the loss gradient; the Hessian bound is on h, not on the training loss. The following gives the particular exact factor and shared clock used here, including interpolation and half-open event semantics.

Lemma F.1 (Original GD block product)**.** *Suppose an actual full-batch GD segment remains in the fixed sample/output-sign domain and has $\gamma_n = \eta_n r_{1,n} > 0$, $2\gamma_n^2 < 1$. Set $P_0 = 1, Q_0 = 0$ and*

$$P_{n+1} = P_n + 2\gamma_n Q_n, \qquad Q_{n+1} = Q_n + \gamma_n P_n.$$

Then

$$u_{j,n}/q_j = P_n h_j - 2Q_n v_j, \quad a_{j,n}/q_j = P_n v_j - Q_n h_j, \quad P_n^2 - 2Q_n^2 = \prod_{k<n}(1 - 2\gamma_k^2) > 0,$$

$P_n > 0$, and $\xi_n = Q_n/P_n$ increases in $[0, 1/\sqrt{2})$. The entire parameter polygon follows the initial line arrangement in increasing clock order. Also

$$\chi_{j,n+1} = (1 - 2\gamma_n^2)\chi_{j,n}.$$

Proof. The actual hidden update matrix is $M(\gamma) = \left(\begin{smallmatrix} 1 & -2\gamma \\ -\gamma & 1 \end{smallmatrix}\right)$. All such matrices commute, since each is affine in one fixed matrix whose square is $2I$. Induction proves the representation. Furthermore

$$P_n \pm \sqrt{2} Q_n = \prod_{k<n}(1 \pm \sqrt{2}\gamma_k) > 0.$$

This gives the determinant and positivity. Division yields

$$\xi_{n+1} = \frac{\xi_n + \gamma_n}{1 + 2\gamma_n \xi_n}, \qquad \xi_{n+1} - \xi_n = \frac{\gamma_n(1 - 2\xi_n^2)}{1 + 2\gamma_n \xi_n} > 0.$$

At parameter fraction $\lambda \in [0,1]$ within one step,

$$\xi_n(\lambda) = \frac{Q_n + \lambda\gamma_n P_n}{P_n + 2\lambda\gamma_n Q_n}, \qquad \xi_n'(\lambda) = \frac{\gamma_n(P_n^2 - 2Q_n^2)}{(P_n + 2\lambda\gamma_n Q_n)^2} > 0.$$

The multiplicative height factor is common and positive. Thus every pair crossing is at its original line clock; at a grid-node collision the two one-sided orientations agree. Finally expanding $2(a - \gamma u)^2 - (u - 2\gamma a)^2 = (1 - 2\gamma^2)(2a^2 - u^2)$ proves the defect formula. A partial segment replaces γ by $\lambda\gamma$. The output bias is still updated by its actual gradient; it has not been removed from the residual. □

A node can lie on a collision wall and lack a strict permutation. The half-open convention assigns that event once to the preceding step. The exact no-skip condition is $\#\{\xi_{ij} \in (\xi_n, \xi_{n+1}]\} \le 1$. Neither this property nor residual-cone safety follows from subcriticality alone; the direct first-exit and clock bounds in the common-trajectory proof establish both for the constructed schedules. In particular, the shared clock does not equate the entire original GF and GD parameter paths at the same physical time.

G A COMPLETE HOMOGENEOUS-INVARIANT LP CRITERION

Outer-sum order realizability has a linear-inequality formulation (Mallows and Vanderbei, 2015, Section 3). Here we also require a safe realization by the original optimizer. Here both output signs are allowed, with $a_j \neq 0$, while $w_j > 0$, $0 < t_j < 1$ and the strict residual cone are retained. This is a different word class from the all-pairs reversal problem: a complete deletion history lists all initially approaching positive–negative pairs and no others.

Theorem G.1 (Homogeneous positive invariant)**.** *Fix $\chi > 0$, a prescribed output-sign pattern, and targets y_0, y_1. A safe strict-cone initialization with $\chi_j = \chi$ exists if and only if there is a positive-output neuron or $y_1 < y_0$. Conditional on this feasibility, a specified generic complete deletion history is realizable by original GF before terminal walls if and only if its transformed initial orders and chronological pair comparisons admit the strict outer-sum linear system described below. The system is equivalent to a unit-margin LP and has an exact Gordan infeasibility certificate.*

Proof. We first translate the event constraints into linear inequalities. We then construct a safe center and place a sufficiently compressed copy of any feasible witness around it. The compression ensures that the optimizer reaches every requested event before any terminal wall.

Let $h = (1-t)/(1+t) > 0$ and $\epsilon_j = \operatorname{sign}(a_j)$. The invariant gives $a_j/q_j = \epsilon_j\sqrt{(h_j^2 + \chi)/2}$ and therefore

$$h_j' = -\epsilon_j\sqrt{2}\sqrt{h_j^2 + \chi}, \qquad H_\chi(t) = -\frac{1}{\sqrt{2}}\operatorname{arsinh}\frac{1-t}{(1+t)\sqrt{\chi}}, \qquad \frac{dH_\chi(t_j)}{d\tau} = \epsilon_j.$$

H_χ is strictly increasing on $(0,1)$. Same-sign order is invariant. Write the initial transformed positions of the positive units as π_i and those of negative units as ν_j, each internally increasing. An approaching pair has $\pi_i < \nu_j$, with unique clock $(\nu_j - \pi_i)/2$. The initial merged order supplies strict linear inequalities in these variables; each consecutive comparison of the desired pair clocks is also linear. Equivalently, reverse the positive index and put $X_r = -\pi_{p+1-r}$, $Y_j = \nu_j$, so clocks are outer sums $X_r + Y_j$ with signed future/complementary constraints. This proves necessity and defines the finite strict system.

For cone feasibility, with all outputs negative the network is nonincreasing, whereas the cone requires $f(1) > y_1$ and $f(0) < y_0$. Hence $y_1 < y_0$ is necessary. For sufficiency choose a common knot $t_* = 1/2$ with $h_* = (1-t_*)/(1+t_*)$ and arbitrary positive scales q_j. Set

$$w_j = q_j(1+h_*)/2, \quad b_j = q_j(h_* - 1)/2, \quad a_j = \epsilon_j q_j\sqrt{(h_*^2 + \chi)/2}.$$

The difference $D = f(1) - f(0)$ is a positive common coefficient times $\sum_j \epsilon_j q_j^2$. With a positive unit, scale it to make $D > y_1 - y_0$; with all negative units and $y_1 < y_0$, choose scales sufficiently small. Then choose $c \in (y_1 - D, y_0)$. The same necessity and a constant predictor handle an empty neuron family.

Given a feasible linear witness, place all its transformed positions at $H_\chi(t_*) + \alpha z_j$, where $\alpha > 0$ is small, and keep the chosen q_j. Invert H_χ to obtain knots; recover w, b, a by the same formulas with the corresponding h_j. These parameters preserve χ, signs and q, and converge to the safe center. Smooth local existence and continuous dependence provide a common safe physical interval and a positive lower bound on r_1 (Teschl, 2012, Theorems 2.2 and 2.9, pp. 38 and 44). Every requested residual clock is $O(\alpha)$, so all are reached in that interval for sufficiently small α. The explicit transformed velocities give every listed crossing, no additional one, and transversality. This proves sufficiency without assuming an unbounded residual clock.

Write the finite inequalities as $Bz > 0$. Dividing any feasible z by its smallest positive row value yields $Bz \geq \mathbf{1}$; the reverse is immediate. Gordan's alternative, the case $A = -B$, $b = 0$ of Boyd and Vandenberghe (2004, Example 2.21), gives infeasibility exactly when there is $\lambda \geq 0$, $\lambda \neq 0$, with $B^\top \lambda = 0$. For rational input matrices the primal and normalized dual witnesses may be chosen rational. A prefix version introduces an observation-clock variable strictly between the listed and unlisted future clocks. This is a feasible subfamily, not a classification of all tractable invariant constraints. □

The exact GD clock proves the corresponding *exclusion*: a safe positive-subcritical GD path starting in a common positive-χ family cannot violate its initial outer-sum ordering or create a same-sign collision, even though its raw balance values may drift. This statement does not assert equal maximal GF and GD prefixes from every fixed initialization.

H Inverse complexity and all quantifiers

For the complexity statement fix positive rational tolerances r, δ, T_0 and, for GD, a positive rational step cap. Choose rational $\varepsilon(m)$ satisfying (13) by a polynomial-time formula, and use the central rational empirical target of the open box. The input is an explicit pair list: it starts in order $1, \ldots, m$, has $N = \binom{m}{2}$ adjacent swaps, uses each pair once and ends in reverse order. Invalid lists are rejected combinatorially.

Theorem H.1 (Fixed-empirical-state inverse completeness)**.** *The existence of a safe original-GF realization of this list, at distance less than r from the specified center, with $q_j = 7$, the fixed empirical moments, $0 < \chi_j - 1 < \delta$, and completion before T_0, is $\exists\mathbb{R}$-complete. The same holds for original GD when one constant rational rate below the cap and any finite number of steps are existentially chosen, with total duration below T_0, all stated safety conditions, and at most one event per half-open step.*

Proof. The reduction uses only the encoded pair list. Its two directions follow from the line representation and the realization theorem; membership is then expressed by a finite geometric formula, without encoding an entire optimizer trajectory.

Hoffmann and Merckx (2018, Theorem 6) establish $\exists\mathbb{R}$-completeness for realizability of *simple allowable sequences*, retaining the full swap order. A point (d_j, C_j) projected in direction $(z, 1)$ has order $C_j + d_j z$. A half-turn can be cut at an event-free direction and represented by a vertical affine sweep; all its finitely many pair intersections lie in a compact interval. Conversely a complete reversal affine sweep includes every pair exactly once and extends to a half-turn without extra crossings. An increasing affine change of sweep coordinate places the compact interval at $[0, 1]$. Thus this geometric problem has exactly the encoded complete-reversal words as inputs, not merely equivalence classes under commuting independent swaps.

Any safe GF realization supplies these lines by the residual representation. Any admissible GD realization supplies them by Appendix F. Conversely, the moment-constrained realization theorem in Appendix C realizes each geometric yes-instance at the prescribed central target; the direct constant-step schedule supplies the discrete realization. The reduction copies the pair list; it does not compute a geometric witness.

For membership, use variables C_j, d_j, s_k with

$$C_1 > \cdots > C_m, \qquad d_1 < \cdots < d_m, \qquad 0 < s_1 < \cdots < s_N < 1,$$
$$C_{i_k} - C_{j_k} + (d_{i_k} - d_{j_k}) s_k = 0 \quad (1 \leq k \leq N).$$

There are $O(m^2)$ variables and constraints of degree at most two. Each pair has distinct slopes and appears once, so every pair crossing is accounted for. The validated combinatorial list supplies adjacency. By the proved bidirectional equivalence, this finite real formula also establishes membership for the optimizer problems. It does not attempt to encode an unknown-length GD path or an ODE in a fixed finite formula. □

Realizable strict geometric words have rational line witnesses by openness of the coefficient inequalities after eliminating intersection times. The exactly moment-matched optimizer initialization may nonetheless be real algebraic, not rational. No polynomial bit-length witness bound in *word* length is established. Nor is hardness established for a preassigned numerical learning rate, a polynomially bounded iteration count, or a fixed arithmetic precision. The resource bound $K = \lceil \max\{40/g_*, 2\varepsilon/(m\eta_{\max})\} \rceil$ is conditional on supplied geometric witnesses and the supplied rate cap; g_*^{-1} may be very large. These restrictions are part of the theorem, not implementation footnotes.

The staircase/sorting-network correspondence in Section 3 of Araujo et al. (2026) sends a sorting network on m labels to staircase shape $(m-1, m-2, \ldots, 1)$. Applying their Theorem 1.3 with $n = m-1$ therefore gives

$$(m-1)^{5(m-1)+o(m)} = m^{5m+o(m)}$$

geometrically realizable complete words. Our main theorem places one optimizer realization of every such word in a single empirical fiber. Counting is imported, and neither a random-initialization probability nor a lower bound on a cell's volume is inferred.

I What a function snapshot does and does not determine

We use breakpoint for the knot location in the BDSO (Breakpoint, Delta-Slope, Orientation) representation of Sahs et al. (2022, Section 2.1). In our positive-slope domain, $t_j = -b_j/w_j$ corresponds to their breakpoint β_i, and the slope jump $a_j w_j$ corresponds to their delta-slope μ_i; the orientation is positive. ReLU identifiability up to positive scaling and permutation has been studied by Stock and Gribonval (2023, Theorems 2,3,5). Here we specify the extra scale or temporal information that recovers raw parameters in the restricted model. This section sharpens the sufficient observation result without claiming a general finite-query algorithm. It also distinguishes the public fixed-q restriction in Theorem C.2 from unrestricted positive parameterizations.

Proposition I.1 (Known-q snapshot recovery)**.** *At a strict-order point in the positive domain, let the initial input function be known exactly on* $[0,1]$*, and let the positive vector* $q_j = w_j - b_j$ *be known in the spatially anchored order. Then the complete original parameter vector is uniquely determined. Writing the ordered knots as* t_j *and their positive slope jumps as* k_j*, the inverse is*

$$c = f(0), \qquad w_j = \frac{q_j}{1+t_j}, \qquad a_j = \frac{k_j(1+t_j)}{q_j}, \qquad b_j = -t_j w_j.$$

Consequently two different histories in the fixed-$q_j = 7$ *family of Theorem C.2 necessarily have different* initial *full input functions. This conclusion holds for GF and for a common specified GD schedule.*

Proof. Since $a_j, w_j > 0$, distributional differentiation gives $\partial_x^2 f = \sum_j k_j \delta_{t_j}$ with $k_j = a_j w_j > 0$. Distinct interior knots therefore determine t_j, k_j, and anchoring assigns the labels. The identity $q_j = w_j(1+t_j)$ yields w_j, then the displayed remaining coordinates. Uniqueness of the original GF or determinism of the specified GD updates proves the last assertion. No temporal branch matching is needed for this fixed-q special case. □

Proposition I.2 (One spatial–temporal first jet recovers unknown scales)**.** *At a strict-order point of the positive two-sample GF domain, suppose the targets are known. The exact functions* $f(x)$ *and* $g(x) = \partial_s f(x)$ *at that time, with one-sided limits of* g *at each knot, determine every original parameter. If* $k_j = [\partial_x f]_{t_j} > 0$*,* $J_j = [g]_{t_j}$ *denotes right minus left limit, and* $r_1 = f(1) - y_1 > 0$*, then*

$$a_j = \sqrt{\frac{-J_j}{r_1(1+t_j)}}, \qquad w_j = \frac{k_j}{a_j}, \qquad b_j = -t_j w_j, \qquad c = f(0). \tag{19}$$

In particular, $J_j < 0$. These observations are idealized full spatial functions and a first temporal derivative, not finitely many noisy test queries.

Proof. Write $f(s,x) = c(s) + \sum_j k_j(s)(x - t_j(s))_+$. On either side of a knot,

$$g(x) = \dot{c} + \sum_j \big(\dot{k}_j(x - t_j)_+ - k_j \dot{t}_j \mathbf{1}_{x>t_j}\big).$$

Its jump is $J_j = -k_j \dot{t}_j$. From the original optimizer, $\dot{t}_j = (a_j/w_j) r_1 (1+t_j)$, so $J_j = -a_j^2 r_1 (1+t_j)$. The positive square-root branch is forced by $a_j > 0$. The remaining coordinates follow. An assigned value of g at the knot itself is immaterial; the one-sided limits suffice. This local formula does not apply at a multiple knot or at $r_1 = 0$ without further analysis. □

I.1 An exact same-initial-function counterexample

Function-preserving scale changes can already alter gradient-flow behavior (Sahs et al., 2022, Section 2.5.1). The example below additionally matches the entire empirical and spectral traces on a common safe interval. The known-q assumption cannot simply be dropped. More strongly, adding the entire empirical and spectral trace need not compensate for the missing individual scales. The following example is separate from the near-balanced, full-language theorem: it has width three, does not fix $q_j = 7$, and compares an empty history with a one-event history.

Proposition I.3 (Same initial full function and diagnostic traces, different safe histories)**.** *With zero targets at inputs $0, 1$ and $c = -1$, initialize the three neurons of two models as follows:*

Model	j	a_j	w_j	b_j
A	*1*	2	1	$-1/4$
A	*2*	$\sqrt{217}$	$\sqrt{217}$	$-\sqrt{217}/2$
A	*3*	$3\sqrt{2}/4$	$24\sqrt{2}$	$-18\sqrt{2}$
B	*1*	2	1	$-1/4$
B	*2*	$\sqrt{434}/4$	$2\sqrt{434}$	$-\sqrt{434}$
B	*3*	6	6	$-9/2$

Their initial functions agree for every real input*:*

$$f_A(0,x) = f_B(0,x) = -1 + 2(x - 1/4)_+ + 217(x - 1/2)_+ + 36(x - 3/4)_+. \tag{20}$$

On a common finite GF interval ending at effective clock $\xi = 1/10$, their entire training prediction, loss, empirical NTK, gradient norm, parameter norm, full Hessian spectrum, and cross-time Euclidean diagnostic traces agree. Model A has no collision on this interval; model B has exactly the event $(1,2)$ at $\xi = 8/91$. At the common final physical time S,

$$f_B(S, 3/5) - f_A(S, 3/5) = \frac{1767}{35} - \frac{279\sqrt{2}}{10} > 10. \tag{21}$$

No population-risk or statistical-generalization assertion is made. The scale and the input are part of this explicit deterministic example.

Proof. We check the example in three parts: equality of the initial functions and empirical traces, physical realization of the two event words, and the final prediction gap. The time-arrival check is necessary even though the candidate collision clocks can be computed exactly.

Both models have ordered knots $(1/4, 1/2, 3/4)$ and positive kink strengths $(2, 217, 36)$, proving (20). Direct sums in the original coordinates give, for both,

$$D = 119, \qquad E = 9137/16, \qquad \sum_j q_j^2 = 64285/16, \qquad r_0 = -1, \quad r_1 = 118.$$

The aggregate ODE therefore has the same initial state. The parameter-norm identity uses only $\sum_j q_j^2$, not the individual vector. In mode coordinates one may extend the fixed isometry by an orthogonal

action on $\bar{q}$ as well: the loss is independent of that mode, and both $\bar{q}$ vectors have the same norm. Together with equal D, E, this proves fixed orthogonal conjugacy of the two safely realized paths, including all stated spectral and cross-time consequences. This is an isometry of the fixed-mask polynomial, not a global function-preserving symmetry along training.

For the effective line coordinates, both have

$$h = (3/5, 1/3, 1/7), \qquad v_A = (8/5, 2/3, 1/56), \quad v_B = (8/5, 1/12, 4/7).$$

The pair clocks, in pair order $(12, 13, 23)$, are

$$(1/7, 64/443, 16/109) \quad \text{for A}, \qquad (8/91, 2/9, -8/41) \quad \text{for B}.$$

Only the first clock of B lies in $(0, 1/10)$. Actual finite-time reachability is not inferred from this list. At $\xi = 1/10$, the common residual test is

$$d = 49/50 > 0, \qquad G = 51419/800 > 0.$$

For model A the three positive values of $q_j^2 d - U_j^2$ are $1127/800, 91791/200, 677799/200$; the values of $a_j A_j$ are $77/20, 4123/20, 9/40$. For B they are $1127/800, 687239/200, 21591/200$ and $77/20, 651/40, 351/10$, respectively. All U_j are also strictly positive. The finite-reachability theorem in Appendix E proves that both actual flows reach the endpoint before any terminal wall. The common aggregate solution makes their residual clocks and physical hitting time S identical. Nonzero pair slopes give the single transverse collision.

At the endpoint, use $u_j = (u_j^0 - 2a_j^0\xi)/\sqrt{d}$, $a_j = (a_j^0 - u_j^0\xi)/\sqrt{d}$, $w_j = (q_j + u_j)/2$ and $b_j = (u_j - q_j)/2$. At $x = 3/5$, only neuron 1 of A is active; neurons 1 and 2 of B are active. Neuron 1 is identical on both paths. The extra contribution of B's neuron 2 simplifies to (21). The bound $\sqrt{2} < 10/7$ gives a lower bound $372/35 > 10$. □

The example is designed to preserve $\sum q_j^2$ by exchanging the two scale branches of neurons 2 and 3. Without that compensating exchange, equal D, E alone would not imply equal total parameter norms. Its initial time-derivative jumps differ, consistently with Proposition I.2. It does not prove that every full reversal can be imposed while also fixing one entire initial function; that stronger occupancy claim is not made.

A bounded-output normalization. Multiply every hidden (a, w, b) coordinate in both models by $1/\sqrt{217}$ and set $c = -1/217$, keeping the targets zero. Then the common initial function is (20) divided by 217; on $[0, 1]$ its range is $[-1/217, 118/217]$. The pair line clocks are unchanged, all endpoint safety signs are unchanged, and G is divided by 217. Reapplying finite reachability gives a common endpoint for the *scaled pair*. The endpoint prediction difference is (21) divided by 217, greater than $10/217$. This is not a claim that uniform parameter scaling preserves the physical time of the unscaled optimizer. It is a separate pair with the same exact diagnostic equalities and a finite nonzero function discrepancy on a bounded initial-output scale.

J Robust multisample separation: complete certificate argument

The three-sample construction requires a different argument from the exact two-sample clock. We start from a pair related by an orthogonal mixing of neurons, then bound independently perturbed executions around that pair. The proof proceeds from parameter tubes to event words, function separation, and observation bands. The numerical inequalities are recorded in a separately archived rational certificate; see `ARTIFACT_MAP.md` for its delivery status. The end-to-end kernel verification of the uncertainty tubes remains incomplete. Outward-rounded enclosures use the general inclusion principle of verified numerics (Rump, 2010, Sections 4 and 5.5); the present optimizer bounds and checker retain their own proof obligations.

J.1 Data and fixed optimization schedule

The original network and weighted half-squared loss are

$$f_\theta(x) = c + \sum_{j=1}^{3} a_j \,\mathrm{ReLU}(w_j x + b_j), \qquad L_d(\theta) = \frac{1}{2}\sum_{i=0}^{2} \lambda_i (f_\theta(x_i) - y_i)^2.$$

Every one of the ten coordinates is trained. Let d_0 have inputs $(0, 1, 2)$, targets $(0, 0, 0)$ and weights $(1, 1, 1)$. The allowed fixed dataset box is

$$|x_i - i| \le 10^{-3}, \qquad |y_i| \le 10^{-3}, \qquad |\lambda_i - 1| \le 10^{-3} \quad (i = 0, 1, 2). \tag{22}$$

All neurons will be inactive at x_0 and active at both x_1, x_2. The two active augmented inputs are not proportional. Their weighted design Gram matrix obeys

$$G_x = \sum_{i=1}^{2} \lambda_i (x_i, 1)^\top (x_i, 1), \qquad \lambda_{\min}(G_x) > 0.1417 > 1/8.$$

Indeed $\det G_x = \lambda_1 \lambda_2 (x_2 - x_1)^2$, and $\lambda_{\min} \ge \det G_x / \operatorname{tr} G_x$ gives the stated bound over (22). None of the samples is duplicated, assigned zero weight, or omitted from the gradient.

Use fixed, nonasymptotic resources

$$\eta = \frac{1}{200} = 0.005, \qquad N = 16, \qquad N\eta = \frac{2}{25} = 0.08. \tag{23}$$

Define initial parameter matrices with rows (a_j, w_j, b_j) by

$$Z_A = \begin{pmatrix} 3/5 & 9/10 & -9/50 \\ 11/10 & 13/10 & -13/25 \\ 4/5 & 17/10 & -221/200 \end{pmatrix}, \quad Q = \frac{1}{27} \begin{pmatrix} 25 & 10 & 2 \\ -10 & 23 & 10 \\ 2 & -10 & 25 \end{pmatrix}, \quad Z_B = Q Z_A,$$

with $c_A = c_B = -1/2$. Explicitly,

$$Z_B = \begin{pmatrix} 46/45 & 389/270 & -397/900 \\ 91/90 & 379/270 & -707/900 \\ 17/45 & 313/270 & -1519/1800 \end{pmatrix}.$$

Direct rational multiplication gives $Q^\top Q = I$. Initial spatial orders in both networks anchor labels as $1, 2, 3$. The nominal A knots are $0.2, 0.4, 0.65$; no coalescing limit is taken.

J.2 Common-mask equivariance

Write $r_i = f_\theta(x_i) - y_i$ and

$$R_0 = \sum_{i=1}^{2} \lambda_i r_i, \qquad R_1 = \sum_{i=1}^{2} \lambda_i r_i x_i.$$

On this fixed activation stratum, the *original* updates are

$$a_j^+ = a_j - \eta(w_j R_1 + b_j R_0), \qquad w_j^+ = w_j - \eta a_j R_1, \tag{24}$$

$$b_j^+ = b_j - \eta a_j R_0, \qquad c^+ = c - \eta \sum_{i=0}^{2} \lambda_i r_i. \tag{25}$$

The corresponding GF uses the same negative gradients. In particular, at d_0,

$$(w_j - b_j)^+ - (w_j - b_j) = -\eta a_j r_2,$$

so the two-sample conserved q_j is no longer conserved. The certificate gives $r_2 > 1$ at every update, even over the wider uncertainty box. At the nominal first two updates the two shared generators do not commute: the determining scalar $R_{1,0} R_{0,1} - R_{0,0} R_{1,1}$ is between -0.611 and -0.610. This is not reuse of the old single-clock integrable system.

Lemma J.1 (Conditional exact empirical equivalence)**.** *For any finite dataset and one group of neurons sharing a fixed sample activation mask, apply the same orthogonal matrix Q to their output-weight, slope and bias vectors, leaving c fixed. If both resulting optimizer paths retain that mask, then the original GF, or GD with common rates, is equivariant under this fixed Euclidean isometry. Its sample predictions, loss, empirical NTK, parameter norm, gradient norm, and full loss-Hessian spectrum agree at corresponding times, nodes, and polygon fractions.*

Proof. For an active sample, the group's contribution is $a^\top(wx_i + b)$, preserved by $a, w, b \mapsto Qa, Qw, Qb$. For an inactive sample its contribution is zero. Thus each fixed-mask polynomial prediction is invariant under an orthogonal linear transformation T of the original parameters. Differentiating gives $\nabla L(T\theta) = T\nabla L(\theta)$ and $\nabla^2 L(T\theta) = T\nabla^2 L(\theta)T^\top$. Uniqueness for GF, induction for GD and linearity for polygon interpolation prove the assertions. The empirical NTK follows by transforming each sample gradient. These statements apply only where both original ReLU masks are valid. □

This symmetry supplies the exact reference pair. It leaves open the essential questions: whether both paths remain safe, whether their events and off-sample functions differ, and whether those differences persist under independent errors. We establish these facts directly for the nonsingular active design, without applying the two-sample realization theorem.

J.3 A UNIVERSAL FINITE-STEP ROBUSTNESS STATEMENT

A perturbed execution is

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n - \eta\nabla L_d(\widehat{\theta}_n) + \zeta_n, \qquad \|\zeta_n\|_\infty \le 10^{-6}. \tag{26}$$

The two executions use the same fixed dataset d in (22); their initialization errors and every ζ_n can be independent and adversarial. No correlation is imposed to preserve the symmetry. Histories refer to the actual piecewise-linear interpolation of these nodes. Node events have the half-open preceding-step convention.

Theorem J.2 (Robust three-sample separation; finite-certificate version)**.** *Use (23) and the rational centers above. For every d in (22), every pair of initializations satisfying*

$$\|\widehat{\theta}^0_A - \theta^0_A\|_\infty \le 10^{-4}, \qquad \|\widehat{\theta}^0_B - \theta^0_B\|_\infty \le 10^{-4},$$

and every error sequence in (26), all of the following hold.

1. *Both complete polygons retain positive a, w and the common sample mask. At time zero both initializations have the anchored left-to-right order $(1, 2, 3)$; their subsequent orders follow the events below. The minimum parameter/activation safety margin is greater than 0.083. The initial adjacent-knot spacing exceeds 0.167.*
2. *A has no collision. B has exactly one oriented polygon crossing of labels $2, 3$, in one of steps 8–15. If it occurs at a node, the two one-sided crossing orientations agree. No step contains multiple events. The event need not occur in the same step for every uncertainty realization.*
3. *At the final node, for every $x \in J = [19/25, 41/50]$,*

   $$f_{\widehat{\theta}^N_B}(x) - f_{\widehat{\theta}^N_A}(x) > 1/25 = 0.04.$$

   Throughout training $|f_{\widehat{\theta}}(x)| < 1.5$ for $x \in [0, 1]$.
4. *Relative to the two corresponding exact symmetry-related runs at the same d, both perturbed runs stay in explicitly bounded observation bands. The raw pair-difference bounds in Table 3 hold on the entire polygons.*

The same conclusions hold with initialization radius 10^{-3} if the gap bound is replaced by $1/40 = 0.025$, the safety margin by 0.074, and the initial spacing by 0.160. The dataset and per-update error budgets and the fixed rate and step count are unchanged.

The quantitative constants in this theorem are certified by the finite rational interval calculation below. The proof is an induction over 16 steps and algebraic inequalities, not extrapolation from randomly tested perturbations.

J.4 STABILITY AND THE TWO ERROR RECURRENCES

A bound using the full Hessian norm in an exponential unnecessarily counts positive curvature as instability. We instead separate the positive-semidefinite Gauss–Newton part. Let $g_i = \nabla f_i$,

Table 3: Deterministic absolute observation bounds for Example 1; these are neither measurements nor relative errors. Units differ by row. Exact symmetry-related runs have zero difference for every permitted dataset. Each perturbed run is within half the displayed bound of their common exact trace.

Raw observable pair difference	init. radius 10^{-4}	init. radius 10^{-3}
Sample prediction, max entry	0.00374	0.03440
Weighted loss	0.01744	0.17554
Empirical NTK, max entry	0.01673	0.15386
Parameter norm	0.001116	0.010182
Gradient norm	0.03002	0.30031
Ordered Hessian eigenvalues, max	0.03354	0.30847

$k_i = \|(x_i, 1)\|$ for active samples, and $k_0 = 0$. Exactly,

$$\nabla^2 L_d = \sum_i \lambda_i g_i g_i^\top + \operatorname{diag}(B, B, B, 0), \quad B = \begin{pmatrix} 0 & R_1 & R_0 \\ R_1 & 0 & 0 \\ R_0 & 0 & 0 \end{pmatrix}.$$

Its repeated residual block has eigenvalues $0, \pm V$ with $V = \sqrt{R_1^2 + R_0^2}$. On any convex safe tube let

$$V \le V_n, \qquad \sum_i \lambda_i \|g_i\|^2 \le A_n.$$

If $\eta A_n \le 2$, then the Euclidean Lipschitz constant of the full GD map is bounded by

$$\|I - \eta \nabla^2 L_d\|_2 \le 1 + \eta V_n. \tag{27}$$

Every Hessian eigenvalue lies between $-V_n$ and $A_n + V_n$. After multiplication by $-\eta$ and addition of 1, the endpoint magnitudes are at most $1 + \eta V_n$ whenever $\eta A_n \le 2$, proving (27). Applying this bound along the segment between two parameters gives a Lipschitz bound for the GD map on the convex safe tube. In the tighter box the certificate has $A_n < 33$ and $\eta A_n < 1$, while $\eta \sum_n V_n < 0.47$. Thus the radius estimate depends on the residual curvature rather than on a full-Hessian exponential $e^{T\|H\|}$.

For data perturbations at a fixed nominal parameter vector, write $S = \sum_j a_j w_j$. For active samples,

$$\partial_{x_i} \nabla L = \lambda_i (S g_i + r_i \partial_{x_i} g_i), \quad \partial_{y_i} \nabla L = -\lambda_i g_i, \quad \partial_{\lambda_i} \nabla L = r_i g_i.$$

The first derivative is zero for the inactive sample. Thus, with $G_i \ge \|g_i\|$ and $G_{x,i} \ge \|\partial_x g_i\|$, a valid data sensitivity bound is

$$C_n = \sum_i \left(\lambda_i G_i + |r_i| G_i\right) + \sum_{i=1}^{2} \lambda_i \left(|S| G_i + |r_i| G_{x,i}\right).$$

All suprema here are over the explicit data box, at the nominal state enclosure. Let $s = 10^{-3}$, $u = 10^{-6}$ and let r_n bound the Euclidean distance from the nominal d_0 trajectory. Starting at $r_0 = \sqrt{10}\rho$, compute an upward-rounded recurrence

$$r_{n+1} \ge (1 + \eta V_n) r_n + \eta C_n s + \sqrt{10} u. \tag{28}$$

For comparison to the exact, same-data symmetry pair, the separate radius satisfies

$$p_0 = \sqrt{10}\rho, \qquad p_{n+1} \ge (1 + \eta V_n) p_n + \sqrt{10} u. \tag{29}$$

The two radii answer different questions. The radius r_n compares to a fixed nominal trajectory, so it must pay for changing the data. It controls masks, collision margins, and the final function gap. By contrast, p_n compares to an exact reference with the very same data. The data-forcing term cancels in that comparison, giving the tighter radius needed for the observation bands. No assumption ties together the initialization or update errors of A and B.

The nominal states are enclosed at 96-bit dyadic precision by the original update equations, using outward rounding and interval inclusion (Rump, 2010, Section 5.5). At step n, a coordinate cube of

radius r_n around the nominal enclosure contains the Euclidean error ball. Interval evaluation first verifies safety on this cube, then bounds V_n, A_n, C_n and checks the radius needed for step $n+1$. This order closes the induction: the stability estimate is used only on a tube whose mask has already been verified. Positive margins at both endpoint states also give polygon safety, because each training preactivation is affine in the segment fraction. The resulting radii satisfy

ρ	r_{16}	p_{16}
10^{-4}	< 0.004327	< 0.000558
10^{-3}	< 0.008883	< 0.005091

No assumption that the perturbed run remains safe is used without closing this induction.

J.5 Certifying the event word, including update errors

Use $F_{ij} = b_i w_j - b_j w_i$. Initial positive values mean $t_i < t_j$. For an exact original step (25), F_{ij} is affine in the interpolation fraction: the quadratic coefficient cancels because the same R_0, R_1 appear for every neuron. This identity does not require two samples or a common scalar residual clock.

Independent update errors remove the exact cancellation. Write $\delta w_j = -\eta a_j R_1 + e_{w,j}$ and $\delta b_j = -\eta a_j R_0 + e_{b,j}$; the segment polynomial is now $F_0 + \lambda F_1 + \lambda^2 F_2$. The quadratic term contains products with at least one error coordinate, so $|e| \le u$ gives

$$|F_2| \le \eta u(|a_i| + |a_j|)(|R_0| + |R_1|) + 2u^2.$$

The difference between this polynomial and its endpoint chord is $F_2\lambda(\lambda - 1)$, whose magnitude is at most $|F_2|/4$. Endpoint signs alone would therefore be insufficient; the certificate includes this correction when checking every segment. All A pairs and B pairs $(1,2), (1,3)$ remain strictly positive, with a smallest wider-box margin above 0.057.

For B pair $(2,3)$, the value at node 7 is positive and that at node 15 is negative throughout the uncertainty sets. On all segments 7 through 14, direct interval evaluation proves

$$\frac{d}{d\lambda} F_{23}(\lambda) < -0.0084$$

in the wider box (less than -0.0093 in the tighter box). All earlier segments are positive and all later segments negative. The endpoint sign change gives existence, and strict decrease throughout the window gives uniqueness and orientation. If the zero is a shared node, the neighboring segments have the same crossing direction and the half-open convention counts it once. Thus the word is established on the actual polygon, including node events.

J.6 Uniform function separation on the test interval

At the nominal final node, B has only neuron 1 active throughout $J = [0.76, 0.82]$. This remains true on the full certified final tubes. With positive output weights, A is convex in the input. Therefore the nominal difference $f_B - f_A$ is concave on J, and its minimum is at an endpoint. Exact interval evaluation of the nominal nodes gives

$$(f_B - f_A)(0.76) > 0.05539569, \qquad (f_B - f_A)(0.82) > 0.05364228.$$

Let M_A, M_B bound the parameter-gradient norms of $f_\theta(x)$ uniformly over the final tubes and $x \in J$. For a potentially active neuron the gradient entries are $(wx + b, ax, a)$; inactive entries are zero, and the bias entry is 1. The same bound applies across a ReLU kink by continuity and piecewise differentiation. Hence

$$(f_{\widehat{B}} - f_{\widehat{A}})(x) \ge 0.05364228 - (M_A + M_B) r_{16}.$$

The complete rational bounds exceed 0.0413502 and 0.0282905, respectively, leaving slack for the reported 0.04 and 0.025 thresholds. The nominal initial functions agree on J, as shown in Appendix K; they are not asserted to agree on the whole input domain.

J.7 Observation bands and a reconstruction lower bound

For fixed data, sample prediction has parameter Lipschitz constant G_i. Loss has constant bounded by $\|\nabla L\|$. Empirical NTK entries have constants at most $k_i G_j + k_j G_i$. Parameter norm is 1-Lipschitz; gradient norm is $(A_n + V_n)$-Lipschitz. Finally,

$$\|\nabla^2 L(\theta) - \nabla^2 L(\widetilde{\theta})\|_2 \le 3 \sum_i \lambda_i k_i G_i \|\theta - \widetilde{\theta}\|_2,$$

because each $\nabla^2 f_i$ is a constant matrix on this stratum. The eigenvalue bound follows from the variational characterization of ordered eigenvalues. Evaluate these constants on hulls of successive nominal nodes, inflated by the larger radius, and multiply by twice the larger p_n. This proves Table 3 on every interpolation fraction.

For the tighter box one may safely replace its six raw pair bounds, in the order in the table, by

$$(0.004,\ 0.018,\ 0.017,\ 0.0012,\ 0.031,\ 0.034).$$

Each execution lies within half these bands of the same exact empirical trace for that dataset. Thus a bounded-error observation model allowing those half-widths may return precisely the same report for both executions. This is a worst-case information statement, not a claim that nonzero differences remain undecodable under infinitely precise observation.

Corollary J.3 (Finite-tolerance reconstruction obstruction)**.** *Let ν be the fixed uniform distribution on $[0, 1]$. Under Theorem J.2's tighter budget,*

$$\|f^N_{\widehat{B}} - f^N_{\widehat{A}}\|^2_{L^2(\nu)} > |J|(0.04)^2 = \frac{3}{31250}.$$

Any reconstruction rule supplied only the common empirical report under the specified bounded-error model has worst-case squared L^2 error at least $3/125000 = 2.4 \cdot 10^{-5}$ on this two-model subclass. The wider budget gives $3/320000$. For randomized rules the same statement holds in expected squared error.

Proof. Integrate the pointwise gap over the fixed length-$3/50$ interval. For any common output h, $\|h - f_A\|^2 + \|h - f_B\|^2 \ge \frac{1}{2}\|f_A - f_B\|^2$. At least one error is at least a quarter of the squared separation; take expectations if the output is randomized. The permitted observation bands admit one common exact-symmetry report. Even with zero observation noise, the exactly related initializations give an obstruction as a subcase. □

This is recovery of the final model from restricted observations, not population prediction risk against a true label function and not an advantage of one training algorithm over another.

K Initial agreement and training-induced interval separation

Let $J = [19/25, 41/50]$. For the exact rational centers of Appendix J, every hidden neuron is initially active on J in both models. Direct rational sums give

$$\sum_j a_j w_j = \frac{333}{100}, \qquad c + \sum_j a_j b_j = -\frac{258}{125}$$

for both models. Thus the eventual gap on J can also be stated as a difference in the functions' training increments.

Proposition K.1 (Matched initially on the eventual test interval)**.** *The nominal three-sample pair satisfies*

$$f^0_A(x) = f^0_B(x) = \frac{333}{100}x - \frac{258}{125} \qquad (x \in J).$$

With independent initialization errors bounded coordinatewise by 10^{-4}, the initial difference obeys

$$\sup_{x \in J} |\widehat{f}^0_B(x) - \widehat{f}^0_A(x)| < 0.0014.$$

Under all the data and update errors allowed in Theorem J.2, the final difference is still greater than 0.04 *on* J*. Hence*

$$(\widehat{f}_B^{16} - \widehat{f}_B^{0})(x) - (\widehat{f}_A^{16} - \widehat{f}_A^{0})(x) > 0.0386 \qquad (x \in J).$$

For the wider 10^{-3} *initialization budget the corresponding initial and final inequalities are* < 0.014 *and* > 0.025*, giving an increment difference greater than* 0.011*.*

Proof. All initial test preactivations remain positive in both coordinate boxes. On that region the original parameter gradient of $f(x)$ has coordinates $(wx + b, ax, a)$ per neuron and a final bias coordinate 1. The segment from a center to its perturbed point remains in the same initial test mask. The mean-value bound in the ℓ^∞/ℓ^1 dual norms gives

$$|f_{\theta+\Delta}(x) - f_\theta(x)| \le \rho \sup_{\vartheta, x \in J} \|\nabla_\theta f_\vartheta(x)\|_1.$$

The two exact rational upper bounds, summed across the two initial boxes, are less than 0.001354588771 for $\rho = 10^{-4}$ and 0.013565543704 for $\rho = 10^{-3}$. Rounding upwards gives the displayed limits. Combine with the separately verified final bounds in Appendix J and subtract the initial difference. □

The initial equality holds on J and a common all-active half-ray. Whole-domain initial equality belongs to the separate two-sample example in Appendix I.1, under different assumptions. The increment bound distinguishes an acquired function difference from an initial discrepancy, but does not attribute that difference causally to the collision.

L Proof dependencies and verification status

L.1 The partial formalization release

The separate V10/V11 development has progressed beyond the historical audit of 20 September 2026. Its frozen release report specifies Lean 4.19.0 (de Moura and Ullrich, 2021), a fixed mathlib revision (mathlib Community, 2020), and recursive audits with leaf axioms restricted to `propext`, `Classical.choice`, and `Quot.sound`. The effective ledger records 261 proved claims, 29 definition alignments, 14 source audits, three partially supported claims, and one unproved external dependency. Among 24 numbered instances in those earlier drafts, 23 are marked proved and one partially supported. Repeated statements are counted separately; these are not completion percentages for the present paper.

The companion digest identifies the committed source, staged audits, and statement names. The latest recorded clean whole-tree build predates later additions. A later source-to-comparison-table audit used source-matched import caches, not a new clean build of the entire tree. Twelve further uncommitted files have only individual build records; two lack confirmed builds. We do not promote these files to audited results. The compact digest supports this status report and the textual correspondence below; the Lean source and build logs remain in separate archives. No new Lean kernel check was performed for this manuscript revision.

L.2 A corrected resource expression

The audit found that an uncapped expression for the selected number of steps in the previous draft's inverse-complexity discussion omitted the learning-rate cap. The corrected formula used throughout this draft is

$$K = \left\lceil \frac{\varepsilon}{m \min\{\varepsilon g_*/(40m), \eta_{\max}/2\}} \right\rceil = \left\lceil \max\left\{ \frac{40}{g_*}, \frac{2\varepsilon}{m\eta_{\max}} \right\} \right\rceil.$$

For $\varepsilon = 1/1000$, $m = 2$, $g_* = 1/2$, and $\eta_{\max} = 10^{-9}$, the selected schedule has $K = 1{,}000{,}000$, whereas the uncapped simplification gives 80. The supplied kernel counterexample establishes this correction. It does not refute the existential realization theorem, and the corrected algebra alone does not prove optimizer completion.

L.3 THE THREE-SAMPLE RESULT AND ITS CERTIFICATES

The current release ledger marks the earlier counterparts of Lemma J.1, Theorem J.2, Corollary J.3, and Proposition K.1 as proved. The reported scope is not limited to the nominal 16-step traces. It includes the shared real data box, independent initialization errors, all allowed update errors, whole-polygon safety and events, both prediction-gap budgets, observation bands, and deterministic and randomized reconstruction bounds. The earlier statement that only the nominal traces had been accepted describes the historical audit, not the scope now reported.

These claims still concern the original optimizer and the specified error model, not arbitrary hardware behavior. The finite-certificate argument requires sound outward arithmetic; correct original gradients on proved masks; valid tube recurrences; whole-segment event checks; and the final function, observation, and reconstruction bounds. A ledger entry does not independently establish these obligations. The compact package does not contain the Lean proof terms or the original three-sample certificate generator and its two budget outputs. It therefore does not replace those reproducibility materials. Numerical experiments do not serve as premises of the universal theorem.

L.4 COMPLEXITY AND EXTERNAL MATHEMATICAL INPUTS

The written proof of Theorem H.1 uses the simple allowable-sequence theorem of Hoffmann and Merckx (2018, Theorem 6). This remains an external mathematical input to the paper. Its absence from the completed Lean dependency chain does not refute the cited theorem. The release reports proved encodings, membership results, and bridges from the geometric language to the original GF/GD realization problems. It also reports a standard-machine polynomial-time compiler from source-formula encodings to the frame-padded comparison table, under the name `sourceComparisonBits_polynomial_time`.

That compiler stops at the comparison table. Certificate search, recursive local replacement, literal history output, and the final hardness composition remain unfinished. Hence the full formal $\exists\mathbb{R}$-hardness chain is not claimed, and the ledger retains partial support for the inverse-completeness statement. No change is made to its mathematical statement or to the existential GD rate and step-count quantifiers.

The staircase correspondence and counting result of Araujo et al. (2026) remain cited inputs to our paper proof. Separately, the release reports proofs of the complete-word and staircase counting bounds. Its unproved historical stronger auxiliary estimate is not used for that reported count. None of these results supplies a typical-initialization probability, a cell-volume bound, or a polynomial bound on the constructed training schedule.

L.5 CORRESPONDENCE WITH THE PRESENT MANUSCRIPT

The report's `thm:main` is the prescribed-state, fixed-$q_j = 7$ theorem, corresponding here to Theorem C.2. It is not the ambient-open parameter statement now numbered Theorem 3.1. The latter also needs Appendix C.4, including the varying-q construction and matching S_q. That additional statement has no supplied release mapping. The table records source-reported coverage, not a new kernel audit of the current types.

Current location	Scope in the supplied release ledger
Appendices B–C, except C.4	Modes, fibers, lift, prescribed-state histories, and relative-open cells: marked proved.
Appendices D–G and I	Function recovery, finite arrival, GD clock, and homogeneous feasibility: marked proved. The approved empty-family support correction is used here.

Current location	Scope in the supplied release ledger (continued)
Appendices J–K	Earlier common-mask, universal three-sample, reconstruction, and increment statements: marked proved.
Theorem H.1	Membership and optimizer bridges are reported proved; full formal hardness remains unfinished.
Theorem 3.1, Appendix C.4	Ambient-open parameter formulation: not covered by the earlier fixed-q theorem alone.
Appendices N–Q	Later planar, switching, and full-rank extensions: no formal coverage supplied by this release.

The source package includes the 24-instance correspondence, exact statement names, the effective claim ledger, and the separate effective and canonical status counts. Source audits and definition alignments are not additional proved theorems. Matching a statement's text does not by itself verify its Lean definitions or hypotheses.

L.6 Separate evidence for the later extensions

The later results retain their paper proofs and stated finite certificates. The switching assumptions in Lemma N.1, the planar and two-hidden-layer bounds, and the full-rank persistence argument are not certified by the older V10/V11 claim count. The two-dimensional and deep experiments likewise supply no new Lean coverage. E1–E3 retain the $1{,}711$ original mixed-unit records; E4–E5 retain 322 separately archived pairs, with the unresolved classifications and checking limits stated in Appendix R. Paper proofs, finite arithmetic certificates, saved-trajectory checks, and kernel-checked declarations remain distinct evidence. The present manuscript is not claimed to have end-to-end formal verification.

M EXPERIMENTAL PROTOCOL, COMPLETE ACCOUNTING, AND NUMERICAL AUDITS

This appendix gives the protocol, complete accounting, and numerical audits of the frozen V10 study. The post-delivery sensitivity checks are reported separately. This appendix preserves the original study and its denominators; the new mathematical extensions and their separate certificate checks are reported in Appendices O–Q, not added to these records.

M.1 EXECUTION AND MEASUREMENT CONVENTIONS

The implemented objective is $\frac{1}{2}\sum_i \omega_i (f_\theta(x_i) - y_i)^2$, with parameter order $(a_1, w_1, b_1, \ldots, a_m, w_m, b_m, c)$; all coordinates are trained on CPU using PyTorch autograd (Paszke et al., 2019), full-batch updates, and no momentum, weight decay, normalization, or symmetry projection. The archived execution used Python 3.11.16, PyTorch 2.7.1, NumPy 2.4.6, SciPy 1.17.1, and mpmath 1.3.0; the frozen environment and source hashes are included in the artifact. NumPy and SciPy are credited to Harris et al. (2020); Virtanen et al. (2020); mpmath and figure-software credits are recorded in `SOFTWARE_CREDITS.bib` supplied with the source. E0 checked the original loss, every gradient coordinate and Hessian at widths $3, 4, 8, 16$, nominal orthogonal centers, event edge cases, the corrected cap regression ($\eta = 5 \times 10^{-10}$, $K = 10^6$, without training a million steps), and historical certificates marked calibration only.

The event checker retains all roots, crossing directions, simultaneous/multiple events, uncertain classifications, and node ownership on $(n, n+1]$; it does not infer events merely from endpoint permutations. For GD the saved binary parameters define the actual polygon; exact rational coefficients and outward root isolation are checked independently, including mask/sign safety and actual data/initialization boxes.

Total update error is evaluated as $\theta^{n+1} - \theta^n + \eta \nabla L(\theta^n)$ at the actual stored data and parameters, using the rational intended learning rate; it includes arithmetic and gradient-evaluation effects as well as injected noise. The original independent Decimal audit covered 1509 pairs and 829760 coordinate updates, with 1666 crossing roots; the acceptance review separately recomputed 5760 coordinates from 18 selected E2 pairs, not all E3 updates again.

For the initial, final, and increment differences we merge all exact binary-rational breakpoints on $J = [19/25, 41/50]$ and $[0, 1]$, evaluate extrema at every piece endpoint, and integrate each affine piece squared exactly; the signed difference is B minus A.

The six diagnostics, in order, are training predictions, loss, empirical NTK, parameter norm, gradient norm, and sorted Hessian spectrum; absolute differences are the largest coordinate/entry differences (scalar differences for norms and loss), while relative differences divide by $\max(1, \|\text{reference A}\|_2)$, using the Frobenius norm for a matrix. Diagnostics are independently evaluated in float64 on saved parameters; GD pair maxima use nodes and segment midpoints, GF comparisons use a common physical-time grid, and accepted GF solver nodes are additionally archived, all as sampled quantities.

A separate exact piecewise-polynomial output audit bounds the saved E2 and nominal polygons by $1.270479 < 1.5$ on $[0, 1]$; this does not turn the six sampled diagnostics or physical-time GF integration into continuous certificates.

M.2 ORIGINAL TASK COUNTS AND COMPLETE FAILURE LEDGER

The original 1711 records contain 1687 executed cases and 24 unrun cases: status counts are 1331 `COMPLETED`, 356 `OUTSIDE_ASSUMPTIONS`, and 24 `RESOURCE_LIMIT_NOT_RUN`. These are mixed record units, not independent sample counts: E1 records are single trajectories, E2/E3 records are network pairs, and stricter GF rechecks are paired numerical repetitions.

The full machine-readable ledger, `data/complete_failure_ledger.csv`, preserves every original record and reason, has 1711 rows and 819350 bytes, and is included unchanged; Table 5 lists every resource-limited record, while the sole E2 exclusion is detailed below.

The remaining 355 outside-assumption records are the 64 wider-radius initialization extensions, 128 non-nominal-rate cases, 160 probe/horizon cases, and 3 controls; their complete factors and observed

Table 4: Complete original accounting. Outside includes intentional extensions and controls; it is not an empirical failure rate. Post-delivery GF sensitivity checks are excluded.

Stage	Planned	Completed	Outside	Unrun
E1_GD	72	48	0	24
E1_GF	120	120	0	0
E1_GF_STRICT	10	10	0	0
E2	1024	1023	1	0
E2_NOMINAL	2	2	0	0
E3_CONTROL	3	0	3	0
E3_INIT	160	96	64	0
E3_PROBE	160	0	160	0
E3_RATE	160	32	128	0

outcomes are reported in Appendix M.5 and in the unchanged ledger. No record was replaced, resampled, moved to a wider box, or removed after inspection; no actual saved run was classified as an unresolved or multi-event GD step, and E0 retains the synthetic edge-case checks.

M.3 E1 CONSTRUCTION, SAMPLING, AND SOLVER SCOPE

E1 implements Theorem C.2's prescribed-state construction, not the specific smaller box displayed in Appendix C.4. It uses widths $3, 4, 6, 8, 12$ for GF, widths $3, 4, 6$ for GD, eight frozen rational line witnesses per width, three target states (center and the two prescribed opposite perturbations), $\epsilon = 1/2000$, and $r = \delta = T_0 = 0.1$; the plus target shifts normalized $(c, D/(49m), E/(49m))$ by $(\epsilon/200, \epsilon/200, -\epsilon/200)$, and minus uses the opposite shift. The witness generator uses NumPy PCG64 (O'Neill, 2014) with the frozen seeds 0–7, cumulative integer intercept/slope increments, distinct exact intersections, and a common centering/scaling; all accepted witnesses were the first candidates, and deduplication changes counts only, not selection. The moment lift is evaluated at 80 decimal digits and then stored in float64; no post-lift symmetry repair is applied. An aggregate reference ODE fixes the common endpoint at residual time $\xi = \epsilon/2$, but each trained GF trajectory integrates the original all-coordinate gradient with SciPy's DOP853 implementation (Virtanen et al., 2020), using relative/absolute tolerances $10^{-10}/10^{-12}$ and maximum step $S/128$; 513 common-time samples, adaptive nodes, and dense interpolation data are saved. DOP853 is an eighth-order explicit Runge–Kutta method with fifth- and third-order error estimators and seventh-order dense output (Hairer et al., 1993, Sections II.5, II.6 and II.10; Appendix, p. 481). The GD schedules use the cap formula in the proof of Theorem C.2; the executed widths $3/4$ use $K = 1647/15011$, and width 6 requires $K = 350009$ and remains unrun.

At a fixed width and target, each trajectory is compared at the same physical times or GD indices against seed 0; Table 7 gives sampled maxima across all such comparisons.

The 10 originally planned tighter-tolerance integrations reproduce the saved nodes because the maximum-step restriction dominates; identical arrays are not evidence of zero integration error.

M.4 E2 FACTORS, DENOMINATORS, AND INDEPENDENT RECHECKS

The 16 cells cross initialization radius $\rho \in \{10^{-4}, 10^{-3}\}$, shared data radius $s \in \{0, 10^{-3}\}$, independent injected coordinate amplitude $u \in \{0, 5 \times 10^{-7}\}$, and float64/float32, with 64 network pairs per cell at $\eta = 1/200$, $N = 16$. Independent PCG64 streams draw each coordinate uniformly from $(-1, 1)$, scaling initialization by ρ and injected noise by u; only data perturbations use the interior factor $(999/1000)s$ before casting to the chosen precision. Within a pair, data are shared and A/B initialization and noise are independent, while corresponding precision runs reuse the same latent arrays and pairing key under the frozen seed derivation.

The eligible counts by (tight64, tight32, wide64, wide32) are $(256, 255, 256, 256)$ out of $(256, 256, 256, 256)$ attempts; the 512 distinct latent cases must not be counted as 1024 independent cross-precision repetitions.

Table 5: All 24 unrun E1 GD records, retained under the original common schedule. These are protocol-enforced resource exclusions, not observed training failures.

Run ID	Width	Target	Seed	K	Frozen cap
E1_GD_00168	6	center	0	350009	200000
E1_GD_00169	6	center	1	350009	200000
E1_GD_00170	6	center	2	350009	200000
E1_GD_00171	6	center	3	350009	200000
E1_GD_00172	6	center	4	350009	200000
E1_GD_00173	6	center	5	350009	200000
E1_GD_00174	6	center	6	350009	200000
E1_GD_00175	6	center	7	350009	200000
E1_GD_00176	6	plus	0	350009	200000
E1_GD_00177	6	plus	1	350009	200000
E1_GD_00178	6	plus	2	350009	200000
E1_GD_00179	6	plus	3	350009	200000
E1_GD_00180	6	plus	4	350009	200000
E1_GD_00181	6	plus	5	350009	200000
E1_GD_00182	6	plus	6	350009	200000
E1_GD_00183	6	plus	7	350009	200000
E1_GD_00184	6	minus	0	350009	200000
E1_GD_00185	6	minus	1	350009	200000
E1_GD_00186	6	minus	2	350009	200000
E1_GD_00187	6	minus	3	350009	200000
E1_GD_00188	6	minus	4	350009	200000
E1_GD_00189	6	minus	5	350009	200000
E1_GD_00190	6	minus	6	350009	200000
E1_GD_00191	6	minus	7	350009	200000

Table 6: E1 complete counts. GF and executed GD words match in every listed run. The 10 original strict-tolerance checks (two center witnesses at each GF width) are additional rechecks, excluded from the GF column. Widths 8 and 12 have no planned GD runs.

Width	Witnesses	Unique words	GF	GD run	GD unrun
3	8	2	24	24	0
4	8	6	24	24	0
6	8	8	24	0	24
8	8	8	24	0	0
12	8	8	24	0	0

Table 7: E1 sampled absolute diagnostic differences. These floating-point measurements do not assert exact equality or a continuous maximum.

Method	f_X	L	K_X	$\|\theta\|$	$\|\nabla L\|$	$\mathrm{spec}(H)$
GF	2.27×10^{-13}	1.34×10^{-11}	9.09×10^{-13}	1.42×10^{-14}	5.91×10^{-12}	1.36×10^{-12}
GD	9.81×10^{-13}	1.82×10^{-11}	4.58×10^{-12}	8.17×10^{-14}	1.45×10^{-11}	5.26×10^{-12}

Within each factor cell the archived Wilson 95% score intervals (Wilson, 1927) use a network pair as the unit, and the percentile-bootstrap intervals for the 5th percentile, median, and 95th percentile use 2000 pair resamples (the percentile construction is described in SciPy community, n.d.; our implementation is custom); `data/cell_statistics.json` retains all of these statistics, including the descriptive all-attempt cell containing the excluded record. The main table instead filters eligibility before taking descriptive extrema and supplies no pooled confidence interval; the per-cell eligible-event interval uses 63 rather than 64 eligible pairs in the affected cell.

Table 8: Every E2 factor cell. Final/increment are minimum signed gaps on J among eligible pairs; initial is the largest eligible $\sup_J$ discrepancy. Total error includes all attempts in the cell. The 63/64 row contains `E2_00425`. Full precision and both eligible/descriptive statistics accompany the source data.

ρ	s	u	Bits	Eligible	Final	Increment	Initial	Total error
10^{-4}	0	0	64	64/64	0.05336	0.05325	5.17×10^{-4}	1.23×10^{-16}
10^{-4}	0	0	32	64/64	0.05336	0.05325	5.17×10^{-4}	7.06×10^{-8}
10^{-4}	0	$5\,10^{-7}$	64	64/64	0.05346	0.05333	4.70×10^{-4}	5.00×10^{-7}
10^{-4}	0	$5\,10^{-7}$	32	63/64	0.05346	0.05333	4.70×10^{-4}	5.39×10^{-7}
10^{-4}	.001	0	64	64/64	0.05341	0.05327	4.41×10^{-4}	1.33×10^{-16}
10^{-4}	.001	0	32	64/64	0.05341	0.05327	4.41×10^{-4}	6.81×10^{-8}
10^{-4}	.001	$5\,10^{-7}$	64	64/64	0.05336	0.05334	3.59×10^{-4}	5.00×10^{-7}
10^{-4}	.001	$5\,10^{-7}$	32	64/64	0.05336	0.05334	3.59×10^{-4}	5.39×10^{-7}
10^{-3}	0	0	64	64/64	0.05200	0.04852	4.81×10^{-3}	1.31×10^{-16}
10^{-3}	0	0	32	64/64	0.05200	0.04852	4.81×10^{-3}	6.80×10^{-8}
10^{-3}	0	$5\,10^{-7}$	64	64/64	0.05112	0.05034	4.15×10^{-3}	5.00×10^{-7}
10^{-3}	0	$5\,10^{-7}$	32	64/64	0.05112	0.05034	4.15×10^{-3}	5.42×10^{-7}
10^{-3}	.001	0	64	64/64	0.05181	0.05007	4.83×10^{-3}	1.29×10^{-16}
10^{-3}	.001	0	32	64/64	0.05181	0.05007	4.83×10^{-3}	6.69×10^{-8}
10^{-3}	.001	$5\,10^{-7}$	64	64/64	0.05113	0.04882	4.61×10^{-3}	5.00×10^{-7}
10^{-3}	.001	$5\,10^{-7}$	32	64/64	0.05113	0.04882	4.61×10^{-3}	5.41×10^{-7}

For `E2_00425` (tight box, float32, $s = 0$, $u = 5 \times 10^{-7}$, seed 31), the actual A initialization error is exactly $1049/10485760 = 0.00010004043579101563$, which exceeds 10^{-4}; B and the data remain in their boxes. This excluded pair nevertheless has the expected event words, final minimum gap 0.05375149, increment minimum 0.05364087, and maximum total update error 5.301×10^{-7}; these observations do not make it tight-box eligible, and it is not reassigned to the wide box.

The acceptance audit recomputed every E2 initial/final/increment function statistic and actual box from raw binary nodes using exact rational arithmetic, and independently checked 98304 quadratic pair-segments across the 1024 pairs; all 1023 eligible pairs had the stated event history. The minimum saved-polygon positive-parameter/original-training activation margin across E2 is 0.0901397951, and all actual update errors are below the 10^{-6} budget.

M.5 E3 SCOPE EXTENSIONS, PROBES, AND CONTROLS

Each initialization, rate, or probe condition uses 32 pairs: initialization scans $\rho = 0, 10^{-4}, 10^{-3}, 0.003, 0.01$ at 16 steps; rates $0.0025, 0.005, 0.01, 0.02, 0.04$ use $32, 16, 8, 4, 2$ steps respectively, keeping duration 0.08.

The probe is $(x, y) = (0.8, 0)$ with weights $0, 0.001, 0.01, 0.1, 1$, at $\eta = 0.005$ for 128 steps and fixed readouts $16/128$; weight zero leaves only the original three effective training samples, although four-input diagnostics still observe the probe location.

All initialization-scan conditions retained distinct words and the original mask/sign domain over their saved windows, including the two radii outside the theorem's prescribed boxes.

All rate cells retained distinct words, but at $\eta = 0.04$ all 32 pairs exited the original domain and the final minimum gaps were negative; $\eta = 0.02$ also fell below the tight-box final-gap reference, outside the fixed-rate theorem's scope.

Table 9: Six *sampled* absolute maxima per E2 cell over nodes and midpoints, using all attempts, including the excluded record. Column units and relative normalization are specified in Appendix M.1. No continuous diagnostic upper bound is inferred.

ρ	s	u	Bits	f_X	L	K_X
10^{-4}	0	0	64	9.97×10^{-4}	5.25×10^{-3}	4.51×10^{-3}
10^{-4}	0	0	32	9.97×10^{-4}	5.25×10^{-3}	4.51×10^{-3}
10^{-4}	0	$5\,10^{-7}$	64	9.88×10^{-4}	5.19×10^{-3}	3.91×10^{-3}
10^{-4}	0	$5\,10^{-7}$	32	9.88×10^{-4}	5.19×10^{-3}	3.91×10^{-3}
10^{-4}	.001	0	64	9.52×10^{-4}	4.96×10^{-3}	3.89×10^{-3}
10^{-4}	.001	0	32	9.52×10^{-4}	4.96×10^{-3}	3.89×10^{-3}
10^{-4}	.001	$5\,10^{-7}$	64	9.37×10^{-4}	4.80×10^{-3}	4.06×10^{-3}
10^{-4}	.001	$5\,10^{-7}$	32	9.37×10^{-4}	4.80×10^{-3}	4.06×10^{-3}
10^{-3}	0	0	64	8.91×10^{-3}	4.59×10^{-2}	4.20×10^{-2}
10^{-3}	0	0	32	8.91×10^{-3}	4.59×10^{-2}	4.20×10^{-2}
10^{-3}	0	$5\,10^{-7}$	64	8.81×10^{-3}	4.43×10^{-2}	4.63×10^{-2}
10^{-3}	0	$5\,10^{-7}$	32	8.80×10^{-3}	4.43×10^{-2}	4.63×10^{-2}
10^{-3}	.001	0	64	8.40×10^{-3}	4.39×10^{-2}	3.63×10^{-2}
10^{-3}	.001	0	32	8.40×10^{-3}	4.39×10^{-2}	3.63×10^{-2}
10^{-3}	.001	$5\,10^{-7}$	64	9.22×10^{-3}	4.85×10^{-2}	3.55×10^{-2}
10^{-3}	.001	$5\,10^{-7}$	32	9.22×10^{-3}	4.85×10^{-2}	3.55×10^{-2}

ρ	s	u	Bits	$\|\theta\|$	$\|\nabla L\|$	$\mathrm{spec}(H)$
10^{-4}	0	0	64	2.21×10^{-4}	8.74×10^{-3}	8.67×10^{-3}
10^{-4}	0	0	32	2.21×10^{-4}	8.74×10^{-3}	8.67×10^{-3}
10^{-4}	0	$5\,10^{-7}$	64	2.49×10^{-4}	8.34×10^{-3}	7.97×10^{-3}
10^{-4}	0	$5\,10^{-7}$	32	2.49×10^{-4}	8.34×10^{-3}	7.97×10^{-3}
10^{-4}	.001	0	64	2.28×10^{-4}	8.05×10^{-3}	7.53×10^{-3}
10^{-4}	.001	0	32	2.28×10^{-4}	8.05×10^{-3}	7.53×10^{-3}
10^{-4}	.001	$5\,10^{-7}$	64	2.16×10^{-4}	7.48×10^{-3}	7.80×10^{-3}
10^{-4}	.001	$5\,10^{-7}$	32	2.16×10^{-4}	7.48×10^{-3}	7.79×10^{-3}
10^{-3}	0	0	64	2.79×10^{-3}	7.58×10^{-2}	7.93×10^{-2}
10^{-3}	0	0	32	2.79×10^{-3}	7.58×10^{-2}	7.93×10^{-2}
10^{-3}	0	$5\,10^{-7}$	64	2.28×10^{-3}	7.14×10^{-2}	7.72×10^{-2}
10^{-3}	0	$5\,10^{-7}$	32	2.28×10^{-3}	7.14×10^{-2}	7.72×10^{-2}
10^{-3}	.001	0	64	1.77×10^{-3}	7.25×10^{-2}	6.95×10^{-2}
10^{-3}	.001	0	32	1.77×10^{-3}	7.25×10^{-2}	6.95×10^{-2}
10^{-3}	.001	$5\,10^{-7}$	64	2.04×10^{-3}	7.72×10^{-2}	7.17×10^{-2}
10^{-3}	.001	$5\,10^{-7}$	32	2.04×10^{-3}	7.72×10^{-2}	7.17×10^{-2}

Table 10: Sampled absolute/relative maxima over eligible pairs in each box; relative values normalize each sampled comparison before taking its maximum.

Box	Scale	f_X	L	K_X
Tight	Abs.	9.97×10^{-4}	5.25×10^{-3}	4.51×10^{-3}
Tight	Rel.	2.08×10^{-4}	4.57×10^{-4}	1.69×10^{-4}
Wide	Abs.	9.22×10^{-3}	4.85×10^{-2}	4.63×10^{-2}
Wide	Rel.	1.93×10^{-3}	4.24×10^{-3}	1.53×10^{-3}

Box	Scale	$\|\theta\|$	$\|\nabla L\|$	$\mathrm{spec}(H)$
Tight	Abs.	2.49×10^{-4}	8.74×10^{-3}	8.67×10^{-3}
Tight	Rel.	8.12×10^{-5}	3.41×10^{-4}	1.75×10^{-4}
Wide	Abs.	2.79×10^{-3}	7.72×10^{-2}	7.93×10^{-2}
Wide	Rel.	1.06×10^{-3}	3.02×10^{-3}	1.61×10^{-3}

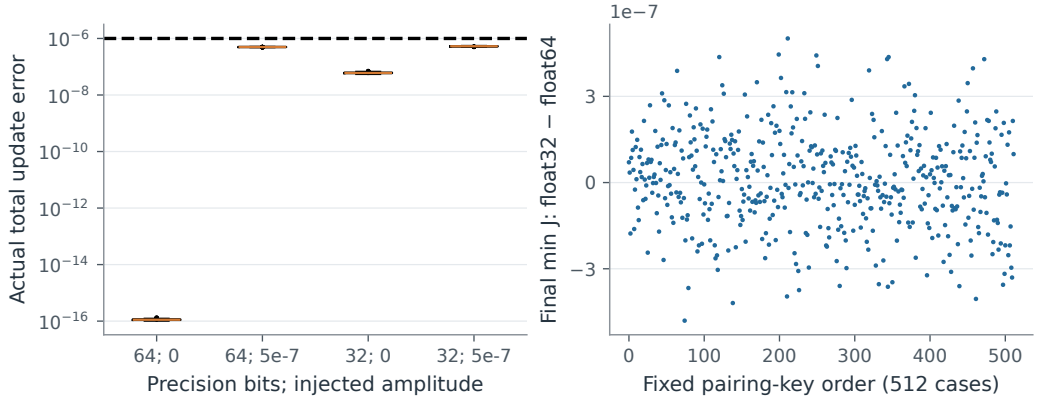


Figure 5: Actual total update errors across all E2 attempts (left; the dashed budget is 10^{-6}), and signed final minimum-gap differences (float32 minus float64) within all 512 matched precision cases (right). Both include the out-of-box record. Initial/noise latents and event words agree across every precision pairing; this dependence precludes treating the two precisions as independent repeats.

Table 11: Complete E3 factor/readout table. Safe means no positive-parameter/original-mask exit through the readout; different words compares anchored histories through that readout. Probe rows at 16 and 128 refer to the same pairs, not additional runs.

Scan	Factor	Step	Pairs	Min. gap	Median gap	Different words	Safe
INIT	0	16	32	0.05350	0.05364	32/32	32/32
INIT	0.0001	16	32	0.05343	0.05364	32/32	32/32
INIT	0.001	16	32	0.05152	0.05355	32/32	32/32
INIT	0.003	16	32	0.04949	0.05359	32/32	32/32
INIT	0.01	16	32	0.03915	0.05155	32/32	32/32
RATE	0.0025	32	32	0.05182	0.05205	32/32	32/32
RATE	0.005	16	32	0.05340	0.05362	32/32	32/32
RATE	0.01	8	32	0.04953	0.04980	32/32	32/32
RATE	0.02	4	32	0.03678	0.03706	32/32	32/32
RATE	0.04	2	32	-0.00547	-0.00518	32/32	0/32
PROBE	0	16	32	0.05346	0.05364	32/32	32/32
PROBE	0	128	32	0.00309	0.00344	0/32	0/32
PROBE	0.001	16	32	0.05345	0.05363	32/32	32/32
PROBE	0.001	128	32	0.00309	0.00331	0/32	0/32
PROBE	0.01	16	32	0.05336	0.05354	32/32	32/32
PROBE	0.01	128	32	0.00289	0.00326	0/32	0/32
PROBE	0.1	16	32	0.05245	0.05263	32/32	32/32
PROBE	0.1	128	32	0.00246	0.00256	0/32	0/32
PROBE	1	16	32	0.04369	0.04385	32/32	32/32
PROBE	1	128	32	0.00316	0.00322	0/32	0/32

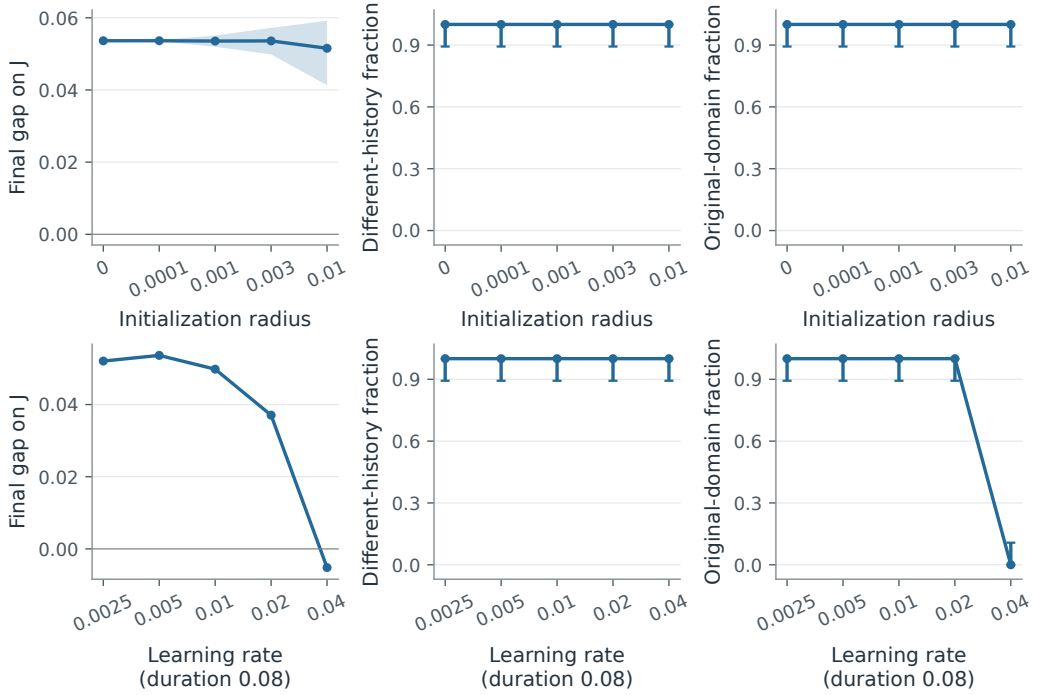


Figure 6: All frozen initialization/rate cells, without additional search. Shading shows empirical 5th–95th percentiles, not confidence intervals; proportion bars are pair-level Wilson 95% intervals. Mask/sign exit at the largest rate limits the interpretation of the short-window observation.

At step 16, each probe-weight cell had $32/32$ distinct histories and preserved the original mask/sign domain; at step 128, each had $0/32$ distinct histories and $0/32$ safe pairs, including weight zero. Consequently the long-horizon change cannot be attributed solely to adding a training point, and an initial history difference does not imply persistent separation of complete long-horizon words.

Identity and neuron-permutation controls have exactly zero initial and final function differences under the saved-parameter rational checks and the same anchored event words; the nonorthogonal diagonal transform is already empirically different at initialization and tests no orthogonal equivariance claim.

Table 12: Deterministic controls, one pair each, excluded from random-pair intervals. The permutation spectrum discrepancy is numerical roundoff; the nonorthogonal transform is a diagnostic control, not a theorem-eligible perturbation. Full-domain exact function metrics are retained in the source JSON.

Transform	Initial $\sup_J$	Final $\sup_J$	Different words	Sampled spectrum
identity	0	0	False	0
swap_neurons_1_2	0	0	False	3.20×10^{-14}
diagonal_1.1_1_1	7.03×10^{-2}	8.22×10^{-3}	False	1.75×10^{0}

M.6 Post-delivery GF numerical sensitivity audit

The acceptance review re-integrated the same 10 fixed center witnesses (seeds $0/1$ at five widths) 30 times with an independent original-coordinate gradient: SciPy DOP853 (Virtanen et al., 2020) at maximum steps $S/256$, $S/512$, and RK45 at $S/512$. The DOP853 method and its dense output are described by Hairer et al. (1993, Section II.6, pp. 194–195; Appendix, p. 481); these method references do not by themselves certify the numerical event times.

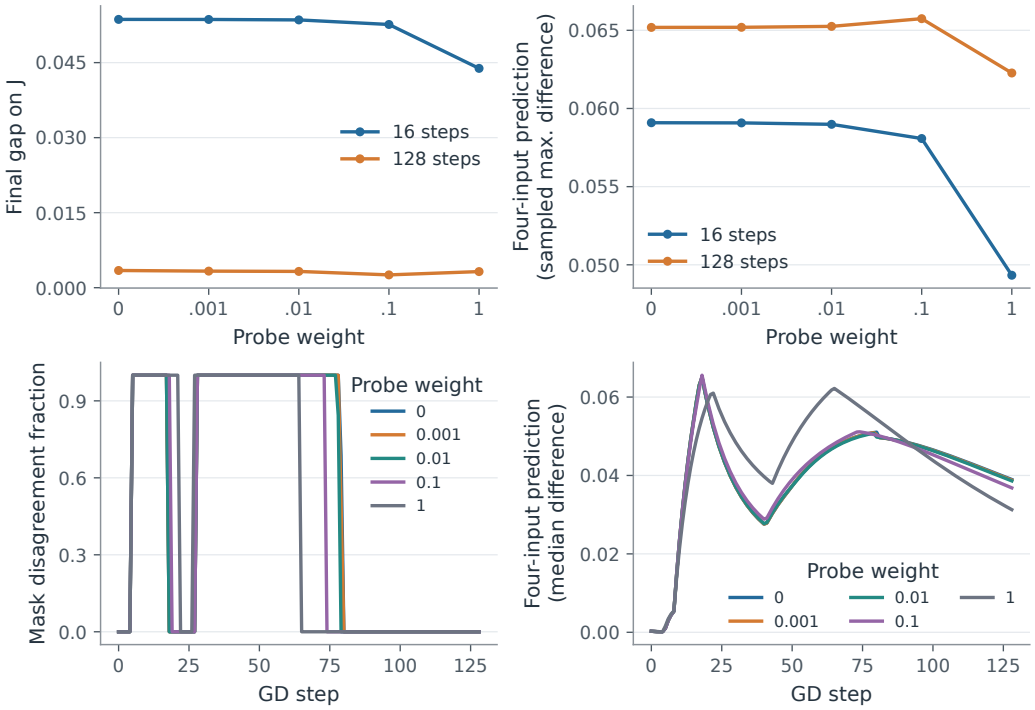


Figure 7: Every probe condition and both fixed readouts. Upper panels show final signed gaps (median and empirical 5th–95th percentiles) and sampled four-input prediction differences. Lower panels show instantaneous A/B mask disagreement and median four-input prediction differences, distinct from the original-domain exit criterion in the factor table.

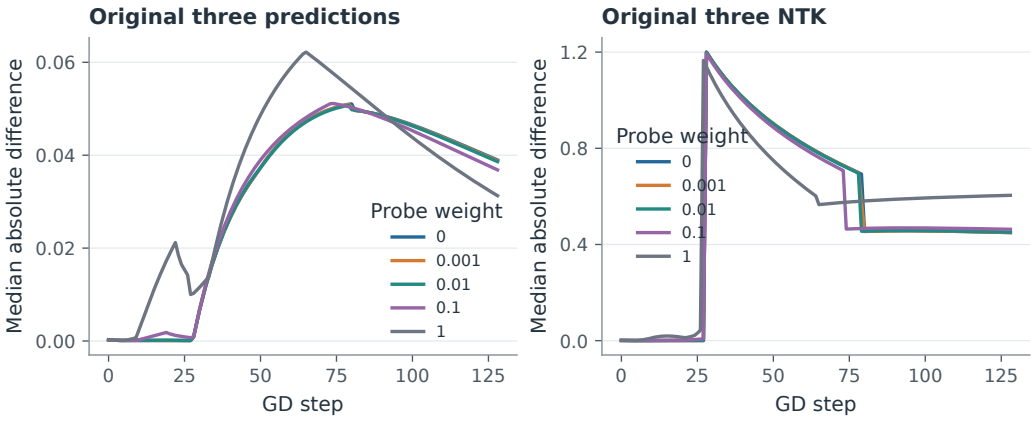


Figure 8: Probe experiments viewed only at the original three training inputs: nodewise median maximum-entry prediction and NTK differences, with the same 32 pairs per weight. These curves are sampled; the added observed location is omitted here, even in the zero-weight ablation.

All event words agreed, with maximum common-grid parameter discrepancy approximately 1.42×10^{-14}. The original `REVIEW_REQUIRED` status remains: two binary64 checks of width 12, seed 1 exceeded the frozen event-time threshold of 10% of the saved minimum event/endpoint gap.

For `E1_GF_00097`, original-gradient RK4/Hermite follow-ups at $(80 \text{ digits}, 256 \text{ steps})$, $(80, 512)$, and $(120, 512)$ agree on all 66 event orders. The high-precision minimum gap is approximately $6.47243398 \times 10^{-11}$, the step-refinement discrepancy is approximately 6.73×10^{-26}, and the original saved times differ by at most about 1.31×10^{-13}.

Exact rational line-clock comparisons and all strict F endpoint tests at $\xi = 1/4000$ also agree for the 120 GF binary initial states, conditional on the paper's line/F reduction.

These $30 + 3$ post-delivery checks remain separate from the original 1711 records and add no independent instances to the 120 GF constructions. High-precision agreement and the rational-clock audit provide no rigorous physical-time ODE error certificate.

Table 13: All 30 post-delivery binary64 sensitivity checks. ID suffix refers to the original E1 GF record. Gap ratio is the event-time discrepancy divided by its saved minimum event/endpoint gap; REVIEW flags above 0.1 are retained. These rows are numerical rechecks, not new preregistered experiments.

Original ID suffix	Method	$S/h_{\max}$	$\max \lvert\Delta\theta\rvert$	$\max \lvert\Delta t\rvert$	Gap ratio	Flag
00000	DOP853	256	7.11×10^{-15}	1.76×10^{-13}	1.08e-07	OK
00000	DOP853	512	1.33×10^{-14}	3.80×10^{-13}	2.34e-07	OK
00000	RK45	512	1.33×10^{-14}	3.80×10^{-13}	2.34e-07	OK
00001	DOP853	256	7.11×10^{-15}	2.73×10^{-13}	2.68e-07	OK
00001	DOP853	512	7.99×10^{-15}	6.63×10^{-14}	6.51e-08	OK
00001	RK45	512	7.99×10^{-15}	6.34×10^{-14}	6.23e-08	OK
00024	DOP853	256	7.99×10^{-15}	1.23×10^{-12}	2.09e-05	OK
00024	DOP853	512	9.77×10^{-15}	3.53×10^{-13}	6.02e-06	OK
00024	RK45	512	9.77×10^{-15}	3.53×10^{-13}	6.02e-06	OK
00025	DOP853	256	1.07×10^{-14}	8.65×10^{-13}	2.12e-06	OK
00025	DOP853	512	1.24×10^{-14}	5.13×10^{-13}	1.26e-06	OK
00025	RK45	512	1.24×10^{-14}	5.13×10^{-13}	1.26e-06	OK
00048	DOP853	256	1.33×10^{-14}	2.40×10^{-12}	0.000152	OK
00048	DOP853	512	9.77×10^{-15}	2.37×10^{-12}	0.000151	OK
00048	RK45	512	9.77×10^{-15}	2.44×10^{-12}	0.000155	OK
00049	DOP853	256	7.11×10^{-15}	6.17×10^{-13}	9.93e-05	OK
00049	DOP853	512	1.15×10^{-14}	6.78×10^{-13}	0.000109	OK
00049	RK45	512	1.15×10^{-14}	6.78×10^{-13}	0.000109	OK
00072	DOP853	256	8.88×10^{-15}	8.65×10^{-12}	0.00855	OK
00072	DOP853	512	1.15×10^{-14}	1.95×10^{-11}	0.0192	OK
00072	RK45	512	1.15×10^{-14}	1.95×10^{-11}	0.0193	OK
00073	DOP853	256	1.33×10^{-14}	3.54×10^{-13}	4.92e-05	OK
00073	DOP853	512	1.42×10^{-14}	1.67×10^{-13}	2.33e-05	OK
00073	RK45	512	1.42×10^{-14}	1.67×10^{-13}	2.33e-05	OK
00096	DOP853	256	9.77×10^{-15}	3.23×10^{-13}	0.000483	OK
00096	DOP853	512	1.42×10^{-14}	5.39×10^{-13}	0.000806	OK
00096	RK45	512	1.42×10^{-14}	5.39×10^{-13}	0.000806	OK
00097	DOP853	256	1.15×10^{-14}	2.08×10^{-13}	0.00321	OK
00097	DOP853	512	1.15×10^{-14}	1.19×10^{-11}	0.184	REVIEW
00097	RK45	512	1.15×10^{-14}	1.19×10^{-11}	0.184	REVIEW

Artifact and evidence boundaries. The artifact retains the complete ledger, frozen protocol, figure data and scripts, unchanged acceptance-review outputs and binary64 warning, sentence-level `PAPER_EVIDENCE_MAP.json`, compilation/rendering records, and a semantic comparison against V10. Measurements, exact checks of saved finite objects, paper proofs, and Lean kernel coverage remain separate. These archived experiments do not enlarge the formalization coverage

Table 14: Post-delivery high-precision follow-up of the flagged record; time differences use the 120-digit, 512-step result as reference. Displayed zero is numerical agreement at the reported precision, not an exact error certificate.

Decimal digits	RK4 steps	Event-order agreement	Time difference
80	256	66/66	6.73×10^{-26}
80	512	66/66	0
120	512	66/66	0

reported in Appendix L; no typical-initialization, general deep-network, cross-wall full-history, population-label-risk, or collision-causality claim follows from them. The separate two-hidden-layer theorem establishes only the specific switching slice-history statement explicitly stated there.

N DEFINITIONS AND THE COMMON PROOF MECHANISM

The data are $d = \{(x_i, y_i, \lambda_i)\}$, with $\lambda_i > 0$, and

$$L_d(\theta) = \tfrac{1}{2} \sum_i \lambda_i (f_\theta(x_i) - y_i)^2, \qquad \theta_{n+1} = \theta_n - \eta \nabla L_d(\theta_n) + \zeta_n.$$

A polygon is the linear interpolation of successive parameter nodes, at physical time $t = \eta(n + s)$, $s \in [0, 1]$. The original gradient is recomputed at every actual node. The additive error ζ_n is a total coordinate update error, not merely an injected random perturbation. The shared dataset is fixed throughout a run; the two runs' initial and update errors need not be related.

The six observations remain those in (2):

$$\mathcal{O}_d(\theta) = \big(f_\theta(X), L_d, K_d, \|\theta\|_2, \|\nabla L_d\|_2, \operatorname{spec}^{\uparrow} \nabla^2 L_d\big), \quad (K_d)_{i\ell} = \langle \nabla f_i, \nabla f_\ell \rangle.$$

Use maximum-entry distance for prediction vectors, kernels and ordered spectra, and absolute distance for scalar quantities. A Hessian at a training wall means a specified one-sided branch value, not a classical second derivative asserted to exist there. All supremum claims below can equivalently exclude the finite walls and include the respective one-sided limits.

A *slice history* records the initial order and the collisions of specified, spatially anchored, genuine kink curves of $s \mapsto f_\theta(s, 0)$ inside a fixed segment. Positive output coefficients and positive effective slice slopes ensure no cancellation. This is not a claim about the complete planar region-adjacency graph. For the deep example the three upper kinks are persistent and individually identifiable on the segment; unrelated lower kinks outside it are not assigned a fictitious label.

GF at training boundaries. The GF clause below concerns existing absolutely continuous paths that solve the original gradient equation almost everywhere and are classical on each open branch interval. Assume finitely many isolated training-gate crossings, all outside the mixed block. At each crossing, one smooth gate $g(\theta) = 0$ changes sign. It satisfies $g(T\theta) = g(\theta)$ near the boundary for the isometry T below, and the adjacent branch fields F_-, F_+ satisfy

$$(\nabla g \cdot F_-)(\theta_*)\,(\nabla g \cdot F_+)(\theta_*) > 0.$$

The continuation immediately follows the branch entered in this common strict direction, with no time spent on the boundary. This convention excludes tangencies, sliding and simultaneous gate changes. It is not an existence or uniqueness claim for arbitrary ReLU gradient flows.

Lemma N.1 (Empirical isometry with specified switching)**.** *For a shallow network $f = c + a^\top [Wx + b]_+$, a common-signature group admits $(a, W, b) \mapsto (Qa, QW, Qb)$ for $Q^\top Q = I$. For a deep network with a common-signature block in its last hidden layer, use*

$$(a_{\mathcal{B}}, V_{\mathcal{B}}, b_{\mathcal{B}}) \mapsto (Qa_{\mathcal{B}}, QV_{\mathcal{B}}, Qb_{\mathcal{B}}).$$

Let T be this fixed isometry, acting identically on the remaining parameters. Take exact runs with $\theta_B(0) = T\theta_A(0)$. If both retain the declared block signatures, then $\theta_B = T\theta_A$ for original GD with common rates and differentiable update nodes, and on its parameter polygons. The same holds for GF under the boundary convention above. All six records agree wherever defined, including matching one-sided limits, and the loss Hessians are orthogonally conjugate. Gates outside the mixed block may switch.

Proof. On an active training input, the block contribution is $a_{\mathcal{B}}^\top (V_{\mathcal{B}} u + b_{\mathcal{B}})$, invariant under T; on an inactive input it is zero. Other contributions are unchanged. Differentiating the polynomial identities on each branch gives

$$\nabla L_d(T\theta) = T \nabla L_d(\theta), \qquad \nabla^2 L_d(T\theta) = T \nabla^2 L_d(\theta) T^\top.$$

The prediction gradients transform in the same way. Orthogonality gives all six record equalities. GD follows by induction, and interpolation commutes with T. For GF, branchwise uniqueness first gives the relation before a crossing. Continuity preserves it at that time. The invariant gate and the two strict normal velocities select corresponding outgoing branches. Uniqueness for those smooth branch fields restarts the relation. Induction over the finite crossings proves the claim. These are conditional statements about the actual ReLU masks and, for GF, the stated continuation; the later finite-step certificates verify their specific GD paths. The argument uses Lemma J.1 and standard symmetry reasoning (Zhao et al., 2023; Grigsby et al., 2023). □

Lemma N.2 (Finite error tubes). *On a verified convex branch tube suppose*

$$0 \preceq \sum_i \lambda_i \nabla f_i \nabla f_i^\top \preceq A_n I, \qquad \Big\| \sum_i \lambda_i r_i \nabla^2 f_i \Big\|_2 \le V_n, \qquad \eta A_n \le 2.$$

If $C_n s_d$ bounds data-induced gradient error, coordinate initialization error is at most ρ, and coordinate update error at most u, valid radii are

$$r_0 = p_0 = \sqrt{p}\rho,$$
$$r_{n+1} = (1 + \eta V_n) r_n + \eta C_n s_d + \sqrt{p}\, u, \tag{30}$$
$$p_{n+1} = (1 + \eta V_n) p_n + \sqrt{p}\, u. \tag{31}$$

Here p is parameter dimension. Radius r_n compares to the nominal-data path; p_n compares to the exact symmetry-related path at the same perturbed data.

Proof. Every loss-Hessian eigenvalue lies in $[-V_n, A_n + V_n]$, so $\|I - \eta \nabla^2 L_d\|_2 \le 1 + \eta V_n$. Apply the mean-value formula on the safe tube, add data forcing and the Euclidean update error, and induct. The data term is absent for the same-data comparison. Safety is checked before this bound is used, then checked at the next inflated endpoint; polygons are checked separately. This is not a full-Hessian exponential estimate or an assumption that perturbed masks remain safe. □

For both models the certificate bounds

$$C_n = \sum_i \big((\lambda_i + |r_i|) G_i + \lambda_i \sum_{k=1}^{2} (|\partial_{x_k} f_i| G_i + |r_i| G_{x_k,i}) \big),$$

where $G_i \ge \|\nabla f_i\|$, $G_{x_k,i} \ge \|\partial_{x_k} \nabla f_i\|$. All suprema include the data box. Smooth observation Lipschitz bounds can be taken as

$$G_{\max}, \quad \sum_i \lambda_i |r_i| G_i, \quad \max_{i,j} (H_i G_j + H_j G_i), \quad 1, \quad A_n + V_n, \quad \sum_i \lambda_i (3 G_i H_i + |r_i| T_i), \tag{32}$$

where $H_i \ge \|\nabla^2 f_i\|_2$, $T_i \ge \|\nabla^3 f_i\|_2$. The last bound and Weyl's inequality control ordered spectra. For shallow branches $T_i = 0$.

N.1 A closed Gram state without a common scalar clock

The following identity explains why the clock is not needed for the selected counterexamples. It generalizes the fixed-signature calculation behind Lemma J.1; it is an application of empirical loss symmetry, not a claim of a new general symmetry principle (Zhao et al., 2023).

Proposition N.3 (Common-signature Gram evolution). *For a shallow network in $\mathbb{R}^d$, assume every neuron in the network has the same training signature $\sigma_i \in \{0, 1\}$ on a safe interval. Write its parameter rows as $Z_j = (a_j, v_j^\top)$ with $v_j = (w_j, b_j)$ and $M = Z^\top Z$. Set*

$$R = \sum_i \lambda_i (f_i - y_i) \sigma_i \bar{x}_i, \qquad B(R) = \begin{pmatrix} 0 & R^\top \\ R & 0 \end{pmatrix}.$$

Then original GF obeys

$$\dot{c} = -\sum_i \lambda_i (f_i - y_i), \quad \dot{Z} = -Z B(R), \quad \dot{M} = -B(R) M - M B(R).$$

The following expressions close this system in (c, M):

$$f_i = c + \sigma_i M_{a,v} \bar{x}_i,$$
$$K_{i\ell} = 1 + \sigma_i \sigma_\ell \left(\bar{x}_i^\top M_{v,v} \bar{x}_\ell + M_{a,a} \bar{x}_i^\top \bar{x}_\ell \right). \tag{33}$$

For original GD, $M^+ = (I - \eta B) M (I - \eta B)$ and $c^+ = c - \eta \sum_i \lambda_i (f_i - y_i)$, with B evaluated at that node. The fractional-step formula holds on every safe parameter segment. Two separately safe trajectories with the same initial (c, M) have the same six empirical observations, including sorted Hessian spectra, at corresponding times. No commutation of $B(R)$ at different times is required.

Proof. Differentiating the common-signature prediction gives $\nabla_{a_j} L = v_j^\top R$, $\nabla_{v_j} L = a_j R$ and the displayed c equation. Multiplication proves the Gram equations. The prediction and kernel formulas follow by summing the individual products, so the empirical system closes. Equal Gram matrices of two m-row factors imply $\widetilde{Z} = QZ$ for some fixed $Q \in O(m)$: the column map preserving their pairwise inner products is an isometry on their column spans and extends orthogonally to $\mathbb{R}^m$. The fixed-signature loss is invariant under this Euclidean parameter transformation. Thus the original safe paths are related by the same Q, and their Hessians are orthogonally conjugate. Also $\|\theta\|^2 = c^2 + \operatorname{tr} M$ and $\|\nabla L\|^2 = (\Lambda r)^\top K(\Lambda r)$. The branch restriction is essential: this is not an invariance of the full-domain ReLU function. □

N.2 Exact check of the mechanism

The planar model uses one inactive and three noncollinear active inputs, so it is not covered by the two-sample reduction. With coordinates of R ordered as $(x, y, 1)$, its first two original nominal updates give

$$R_0 = \left(\frac{5229}{500}, \frac{37}{800}, \frac{2931}{500}\right)^\top,$$

$$R_1 = \left(\frac{357244293174497}{40000000000000}, \frac{22431603458181}{512000000000000}, \frac{395822241103037}{80000000000000}\right)^\top.$$

The first diagonal block of $[B_0, B_1]$ is zero and the lower block is $R_0 R_1^\top - R_1 R_0^\top$. In particular its (x, y) entry equals

$$\frac{11550543703177337}{256000000000000000} > 0.0451,$$

and exact rational arithmetic gives $0.7505 < \|[B_0, B_1]\|_F^2 < 0.7506$. The first changes in $(w_{xj} - b_j)_{j=1}^3$ are

$$\left(-\frac{3447}{250000}, -\frac{12639}{500000}, -\frac{1149}{62500}\right).$$

These calculations rule out reuse of the particular commuting-generator/shared-q_j construction in Appendix F; they do not claim that every possible nonlinear change of clock is impossible. The separation itself is independently established by Theorem O.1 over the full uncertainty boxes.

What this stress test does and does not explain. The scalar clock and line-arrangement lift are crucial to the prescribed-state *full-language proof*; this example does not remove them from that stronger theorem. But they are not necessary for a pair with identical complete empirical traces and different histories. In the latter problem, the operative mechanism is that empirical dynamics factors through a Gram state while individual kink comparisons need not. The deep example further shows that a gate outside a mixed block can change without destroying that block's empirical isometry. It does not show invariance when arbitrary gates inside that block split their signatures, nor typicality under random initialization.

Restriction tested	What the example establishes	What is not established
One input dimension / two samples	The planar pair has noncollinear data, $\lambda_{\min} > 0.14$, and trains both directions.	Full-language realization at each prescribed planar empirical state.
Common scalar generator and conserved $w - b$	Both fail in the exact two-update calculation, but empirical equivalence and distinct histories persist.	Exclusion of every conceivable nonlinear reduction of these trajectories.
No training activation changes anywhere	The deep pair has an effective lower gate change before its distinct upper-kink history.	Invariance after arbitrary events split the mixed block itself.
Selected common-signature block and safe finite window	These remain verified sufficient ingredients of the constructions.	A necessary characterization of all empirical ambiguity, typicality, or long-time persistence.

O A GENUINELY PLANAR FIXED-RATE ROBUST SEPARATION

Use the four nominal inputs, weights and targets

$$X = \big((0,0),(1,-1/2),(1,1/2),(2,0)\big), \quad \lambda = (1,1/2,1/2,1), \quad y = 0. \tag{34}$$

The rows of $Z = (a, w_x, w_y, b)$ are

$$Z_A = \begin{pmatrix} 3/5 & 9/10 & 1/10 & -9/50 \\ 11/10 & 13/10 & 3/20 & -13/25 \\ 4/5 & 17/10 & -1/20 & -221/200 \end{pmatrix}, \quad Z_B = QZ_A,$$

$$Q = \frac{1}{27}\begin{pmatrix} 25 & 10 & 2 \\ -10 & 23 & 10 \\ 2 & -10 & 25 \end{pmatrix}, \qquad c_A = c_B = -1/2. \tag{35}$$

All 13 coordinates train. The first input is inactive for all three neurons, the other inputs active. Their active augmented Gram is nominally

$$G_X = \begin{pmatrix} 5 & 0 & 3 \\ 0 & 1/4 & 0 \\ 3 & 0 & 2 \end{pmatrix}, \qquad \lambda_{\min}(G_X) = \frac{7-\sqrt{45}}{2} > 0.1458.$$

No active input is duplicated, assigned zero weight, or placed on the same affine line as both others.

Theorem O.1 (Planar robustness, all coordinates trained)**.** *Set $\eta = 1/200$, $N = 16$. Perturb each coordinate of the fixed inputs, labels and weights by at most 10^{-4}. Independently perturb each initialization coordinate by at most 10^{-4} and each update coordinate by at most 10^{-6}. For every such pair of runs:*

1. *The common training mask, positive a, w_x, and slice labels $1, 2, 3$ are retained. The active augmented design has $\lambda_{\min} > 7/50$. Along the whole polygons the first/third weight-normal minor has absolute value > 0.177, so the model is not a one-dimensional ridge in disguise. Every original gradient coordinate is nonzero at the update nodes; in particular all second-input-direction coordinates genuinely update.*

2. *On $y = 0$, A has no collision and B has exactly one oriented collision of labels $(2, 3)$ in steps 8–15. Node collisions have this manuscript's preceding-half-open-step convention and agreeing one-sided orientations.*

3. *On the fixed rectangle*

$$J_P = [19/25, 41/50] \times [-1/20, 1/20],$$

$f_B^N - f_A^N > 0.045$. The initial discrepancy is < 0.0014, and the difference of training increments is > 0.046 by the stronger certificate values. On $\Omega = [0,1] \times [-1/2, 1/2]$, $|f| < 1.5$ throughout.

4. *The six full-polygon pair observation bands, in the order defining $\mathcal{O}$, are bounded by*

$$(0.0043,\ 0.021,\ 0.019,\ 0.0013,\ 0.035,\ 0.039).$$

Each run is within half these bands of the common exact same-data reference record. Exact center runs have identical complete observation traces.

Proof with finite arithmetic specification. Apply Lemma N.1 to (35); $Q^\top Q = I$ holds over the rationals. Original gradients at active inputs use $R = \sum_{i=1}^{3} \lambda_i r_i \bar{x}_i$ and are

$$\nabla_{a_j} L = (w_{xj}, w_{yj}, b_j)R, \quad \nabla_{(w_{xj}, w_{yj}, b_j)} L = a_j R, \quad \nabla_c L = \sum_i \lambda_i r_i.$$

The residual Hessian has repeated blocks $\left(\begin{smallmatrix} 0 & R^\top \\ R & 0 \end{smallmatrix}\right)$, hence $V_n = \|R\|_2$. Evaluate (30)–(31) using outward dyadic arithmetic and the original updates. The script `certify_planar.py` verifies every denominator-free safety inequality before taking each update, checks $\eta A_n \le 2$, and checks every endpoint hull. Sylvester's criterion for $G_X - (7/50)I$ holds over the full data box.

For each pair put $F_{ij} = b_i w_{xj} - b_j w_{xi}$. Along an actual noisy step it is quadratic. The code forms all coefficients from the original update and independent additive error intervals. An excluded crossing is ruled out by the smaller endpoint margin minus one quarter of the absolute quadratic coefficient. For B's $(2, 3)$ pair it checks positive sign at node 7, negative sign at node 15, and a strictly negative derivative on all intervening segments; outside this window the sign is strict. All other pairs remain positive. Positive coefficients ensure the curves are genuine kinks, not arbitrary parameter labels.

At the final patch, B has only its first neuron active. Its function is affine there; A is convex because its output coefficients are positive. Thus nominal $f_B - f_A$ is concave, and its minimum is bounded by the four vertices. Parameter-gradient bounds transfer this to the inflated final tubes. The initial patch activates all neurons in both center models, so initial functions agree exactly there. The independent initialization bound uses the ℓ_1 parameter-gradient norm. Formula (32) with $T_i = 0$ and the same-data radius proves the diagnostic bands on every polygon fraction. The complete finite inequalities are in `PLANAR_CERTIFICATE.json`; selected outward-rounded consequences are:

Quantity	Verified strict bound
Maximum final nominal/data radius	< 0.001064
Minimum whole-polygon sign/parameter margin	> 0.05522
Minimum noncollision margin	> 0.04982
B $(2, 3)$ crossing derivative, segment fraction	< -0.01010
Final patch gap after all errors	> 0.04761
Initial discrepancy	< 0.001383
Training-increment gap	> 0.04623
Second-direction gradient magnitude at every update	> 0.00356

These are finite continuum enclosures, not success frequencies of randomly drawn perturbations. Their integer endpoint computations enclose the exact rational centers and original real updates. □

Corollary O.2 (Positive-area restricted reconstruction obstruction)**.** *For $\nu = \mathrm{Unif}(\Omega)$, the area of Ω is one and that of J_P is $3/500$. Hence*

$$\|f_B^N - f_A^N\|_{L^2(\nu)}^2 > \frac{243}{20\,000\,000}.$$

If the observation model permits the half-widths in Theorem O.1, a common exact-reference report is admissible for both runs. Every reconstruction rule has worst-case squared $L^2(\nu)$ error at least $243/80\,000\,000$ on this pair. The expected-squared version holds for randomized rules.

Proof. Integrate the uniform gap over the rectangle and use $\|h - f_A\|^2 + \|h - f_B\|^2 \geq \frac{1}{2}\|f_A - f_B\|^2$. This concerns reconstructing trained functions from restricted observations, not population-label risk or causality of collisions. □

P TWO HIDDEN LAYERS, A REAL TRAINING WALL, AND PERSISTENT AMBIGUITY

The fully connected model is

$$u = \left[P(x, y, 1)^\top\right]_+, \qquad f = c + a^\top \left[Vu + b\right]_+, \quad P \in \mathbb{R}^{2\times 3}, \quad V \in \mathbb{R}^{4\times 2}.$$

There are 23 trainable coordinates. Use data (34), now with coordinate data uncertainty $s_d = 10^{-6}$. Define model B by

$$P = \begin{pmatrix} 1 & 1/5 & 1/10000 \\ 1/5 & 1 & 1 \end{pmatrix}, \quad V_B = \begin{pmatrix} 1 & 1/20 \\ 13/10 & 1/10 \\ 9/10 & 3/20 \\ 1/10 & 1/5 \end{pmatrix},$$

$$a_B = (3/5, 1, 3/10, 1/5)^\top, \qquad c = -1/2.$$

Put $s = (9/20, 5999/10000, 6001/10000)$. If $p_x = P_{:,1}$ and $p_b = P_{:,3}$, set

$$(b_B)_k = -(V_B p_x)_k s_k - (V_B p_b)_k \quad (k = 1, 2, 3), \qquad (b_B)_4 = 1/5. \tag{36}$$

Let $Q_4 = \mathrm{diag}(Q, 1)$ and

$$(V_A, b_A, a_A) = (Q_4^\top V_B, Q_4^\top b_B, Q_4^\top a_B).$$

The lower parameters and c are unchanged. The first three upper units share signature 0111; unit 4 is active at all training inputs. Both lower units are initially active at every training input. The first lower unit deactivates at input zero during the first actual GD segment. The fourth upper unit gives this wall nonzero output sensitivity, approximately 0.02; it is not a blocked-output or zero-gradient gate. It also shares both lower features with the mixed upper block, so the construction is not a direct sum of disconnected subnetworks.

Theorem P.1 (A robust switching two-hidden-layer ambiguity)**.** *For the centers* (36)*, set* $\eta = 1/50000$*,* $N = 128$*, so* $T = 8/3125$*. Allow shared coordinate data errors* 10^{-6}*, independent coordinate initial errors* 10^{-6}*, and independent per-step coordinate update errors* 10^{-8}*. Then every allowed pair of original GD polygons has the following properties.*

1. *The first lower unit at sample zero has exactly one training-wall deactivation, inside*

$$\mathcal{W} = [\eta/2, 17\eta/20] = [10^{-5}, 1.7 \cdot 10^{-5}].$$

   It is strictly decreasing there. No other training mask changes. Every parameter gradient coordinate at every update has magnitude > 0.0300*, and even after the allowed update error every coordinate update is nonzero. The lower input matrix determinant remains* > 0.928*. The three active augmented training inputs have* $\lambda_{\min} > 0.14$*.*

2. *On the fixed probe segment* $(s, 0)$*,* $s \in [1/4, 7/10]$*, both lower gates and upper gate 4 are strictly active. The three other upper units give genuine positive kinks, anchored by their initial order. A has no collision among these kinks. B has exactly one* $(2, 3)$ *collision, in steps 2–4, after the training-wall step.*

3. *On*

$$J_D = [61/100, 64/100] \times [-1/100, 1/100],$$

   the final difference and the difference of training increments are both > 0.022*. The independent initial discrepancy is* $< 2.4 \cdot 10^{-5}$*. Exact initial functions agree on the whole patch. Throughout* Ω*,* $|f| < 1$*.*

4. *Exact same-data reference runs have the same complete empirical and spectral trace, despite the training-wall change and the different kink histories. The independent-error versions have the observation bounds specified in Table 15. In particular the small nodewise bounds are not silently promoted to small synchronous suprema through the wall.*

Proof with finite arithmetic specification. The upper three-unit isometry in Lemma N.1 is valid for either lower mask: sample zero sees none of the mixed block, and the remaining samples see the full mixed block. Thus it intertwines the original update even after the lower mask changes. All coordinates, including those of the untransformed unit, are updated by the full shared residual. No optimizer step retains an obsolete lower mask.

For the finite tubes, f_i is degree three on each branch. With lower mask s and upper mask t, put $d_k = a_k t_k$ and $e_j = s_j \sum_k V_{kj} d_k$. Then the prediction gradient blocks are

$$\nabla_P f = e\bar{x}^\top, \quad \nabla_V f = du^\top, \quad \nabla_b f = d, \quad \nabla_a f = z, \quad \nabla_c f = 1.$$

The nonzero prediction-Hessian blocks are

$$H_{a_k b_k} = t_k, \quad H_{a_k V_{kj}} = t_k u_j, \quad H_{a_k P_{j\ell}} = t_k s_j V_{kj} \bar{x}_\ell, \quad H_{V_{kj} P_{j\ell}} = t_k s_j a_k \bar{x}_\ell.$$

A Frobenius bound for the third derivative is $\sqrt{6(\sum s_j)(\sum t_k)}\,\|\bar{x}\|_2$. Use $V_n = \sum_i \lambda_i |r_i| \|H_i\|_F$, then apply Lemma N.2. The first update is taken on the strictly active source branch. The next node is verified strictly inactive at the one switched coordinate. All later updates use that new branch.

Every upper preactivation is evaluated using the actual interval ReLU on each full parameter hull, so its fixed sign is verified even in the first lower-wall segment. The target lower preactivation is affine in the segment fraction. Its interval value is positive at $1/2$, negative at $17/20$, with strictly negative endpoint slope. This proves a real wall on the actual polygon, not a prescribed path crossing unsupported by GD.

On the probe the lower gates are active, so upper unit k has effective slope $w_k = V_k P_{:,1}$ and intercept $\widetilde{b}_k = b_k + V_k P_{:,3}$. Along a parameter segment these are quadratic polynomials. The pair comparison $\widetilde{b}_i w_j - \widetilde{b}_j w_i$ is quartic. The certificate constructs all coefficients from the original gradient and additive-error intervals, proves all A comparisons and B comparisons $(1,2), (1,3)$ positive, and proves the B $(2,3)$ polynomial strictly decreasing on every segment. It is positive at node 1 and negative at node 4. Thus the unique geometric collision occurs after the training wall. Endpoint-chord corrections include all higher coefficients, so endpoint orders alone are not used as a proof.

At the final patch B has only upper units 1 and 4 active. A has positive upper output coefficients and an affine, all-active lower map on the patch, so A is convex and B is affine. Concavity of $f_B - f_A$ reduces the nominal minimum to four vertices. A bound over all possible A upper masks controls its parameter Lipschitz constant across patch kinks. The final tube correction and initial all-active identity give the claimed gaps.

The executable dyadic certificate gives, without sampling perturbations,

Quantity	Verified bound
Final radius r_N, either model	$< 1.185 \cdot 10^{-5}$
Minimum gradient-coordinate magnitude	> 0.030078
Lower input-normal determinant	> 0.92863
Minimum excluded kink-pair margin	> 0.04759
Training-wall speed magnitude (polygon time)	> 6.9444
Final patch gap after all errors	> 0.0225429
Initial discrepancy	< 0.00002313
Training-increment gap	> 0.0225198

All output coefficients and relevant effective slopes are positive. The same first-hull check bounds both neighboring original GF normal velocities below -6. Thus the boundary has a genuine local GF continuation as well; the 128-step history and gap constants remain GD statements.

The finite inequalities are stored in `DEEP_CERTIFICATE.json`. The generating program is `certify_deep.py`. □

P.1 Correct treatment of discontinuous diagnostics

The two independently perturbed walls need not occur at identical times. At matching branch points, formula (32) and the same-data radius give the small pair band b. Inside $\mathcal{W}$ the norm, loss and predictions remain parameter-Lipschitz, but the other three observations may jump. At the switched sample, let $\beta = a_4 V_{41}$ and $n = \nabla h$; only the fourth upper path is open. The gradient changes by βn. Bounds valid on the first hull are

$$\begin{aligned} J_K &\le \max\{|\beta|\|n\|G_{\max}, \beta^2\|n\|^2\}, \\ J_{\nabla L} &\le \lambda_0 |r_0||\beta|\|n\|, \\ J_{\nabla^2 L} &\le \lambda_0\{2|\beta|G_0\|n\| + \beta^2\|n\|^2 + 2|r_0|\|n\|\sqrt{a_4^2 + V_{41}^2}\}. \end{aligned}$$

The same bounds control the corresponding scalar norm or ordered-spectrum jumps. A safe full pair bound is $b + 2J$. Since $|\mathcal{W}|/T = 7/2560$, the time-average pair bound is $b + (14/2560)J$. This explicitly accounts for asynchronous mask changes.

Proposition P.2 (A synchronous-sup obstruction, not a failure of the separation theorem)**.** *One cannot replace the full-sup column of Table 15 by a bound proportional to vanishing independent errors, even for this example. In particular the nominal total NTK has a nonzero jump at the training wall.*

Table 15: Deterministic pair bands for the same independent-error class. Time average means $T^{-1}\int_0^T |\Delta\mathcal{O}_j(t)|\,\mathrm{d}t$, with the declared entrywise distances. These are different observation metrics, not interchangeable summaries.

Observation	Nodes / outside $\mathcal{W}$	Full synchronous sup	Time-average L^1
Predictions	0.00015	0.00015	0.00015
Loss	0.00043	0.00043	0.00043
Total NTK	0.00320	0.268	0.00393
Parameter norm	0.000024	0.000024	0.000024
Gradient norm	0.00210	0.01891	0.00215
Ordered Hessian spectrum	0.00819	0.474	0.00946

Proof. At the boundary the inactive sample gradient has zero components in the first lower row; the active gradient adds $\beta\bar{x}_0$ there. Thus $|\Delta K_{00}| = \beta^2\|\bar{x}_0\|^2 > 0$. Initially $\beta = 1/50$, and the certified first-step neighborhood retains a strictly nonzero value. Perturb only the first-step additive update to that lower bias by $e > 0$. The gate on that polygon is $h_0 + s(-\eta g_b + e)$, so its crossing time changes strictly with e, although parameter paths converge uniformly as $e \to 0$. At a time between the two walls the kernel difference converges to the nonzero jump. Hence its synchronous supremum does not tend to zero. Example 1 has no such jump and is not contradicted. □

There is also a small aligned-time alternative. Let ψ be increasing, linear on the two pieces of the first step, mapping the A wall to the B wall, with $\psi(0) = 0$, $\psi(\eta) = \eta$, and $\psi(t) = t$ later. The certificate gives

$$\|\psi - \mathrm{id}\|_\infty < 1.397 \cdot 10^{-6}, \qquad \|\theta_A(t) - T^{-1}\theta_B(\psi(t))\|_2 < 3.8791 \cdot 10^{-5}.$$

To verify it, bound the gate-time difference by $2\|\bar{x}_0\|p_{\max}/\gamma$, where $\gamma > 6.9444$ is the first-segment gate speed, then add polygon speed times that time difference to the $2p_{\max}$ same-time parameter bound. A separately verified 0.01 hull around the A first segment contains the aligned parameter segment. Formula (32) gives aligned diagnostic bounds

$$(0.000264, 0.000954, 0.005890, 0.000039, 0.004191, 0.015310).$$

This is explicitly a time-aligned metric, not an unnoticed change to synchronous observations.

Corollary P.3 (Positive-area function obstruction through a real training-wall change)**.** *With $\nu = \mathrm{Unif}(\Omega)$, $|J_D| = 3/5000$ and*

$$\|f_B^N - f_A^N\|_{L^2(\nu)}^2 > \frac{363}{1\,250\,000\,000}.$$

For the exact reference pair, the whole empirical trace is identical, so the noiseless restricted-record reconstruction lower bound is one quarter of this number. The same conclusion for independently perturbed runs uses half of the selected bands in Table 15; it must specify whether the observation contract is nodewise, full synchronous, or time-averaged. The bound is not a population-risk statement or proof that the spatial collision caused the gap.

Q A USEFUL, LIMITED SUPPLEMENT TO THE COMPLETE-HISTORY THEOREM

Theorem 3.1 already covers every finite $m \geq 2$. Merely adding a larger-width simulation is not a theorem extension. The following is a genuine higher-dimensional, full-rank data statement, but deliberately does *not* advertise exact empirical equality under arbitrary designs.

Theorem Q.1 (Full-rank finite-resolution persistence of the history language)**.** *Fix $d \geq 2$, $m \geq 2$, tolerances and ε as in* (13)*, a source state $(c_*, D_*, E_*) \in \mathcal{B}_{m,\varepsilon}$ from* (14)*, and a finite family $\mathcal{W}$ of stretchable complete reversals. For every $\varepsilon_{\mathrm{obs}} > 0$ there exist $d + 2$ distinct inputs in $\mathbb{R}^d$, positive weights bounded below by $1/(d+1)$, zero targets, and one all-parameter shallow-network initialization per word such that:*

1. *The active augmented design has rank $d + 1$. Every transverse input coordinate has a nonzero original gradient, and the initial input-weight matrix has rank $\min(m, d)$; this is not a zero-padding construction.*

2. *On the public probe $x = se_1$, all words are exactly realized by original GF on one common physical interval. Every tracked intersection is a genuine positive function kink, simple and transverse, and no unwanted pair collision occurs.*

3. *On that interval every pair of six-component empirical records differs by less than $\varepsilon_{\mathrm{obs}}$ in the respective fixed raw norms. This remains true, with strict history and safety margins, for sufficiently small independent initialization and shared data perturbations and sufficiently small continuous additive field errors.*

The family may be the full stretchable language at the fixed width, since that language is finite. This does not give an architecture-uniform conditioning bound, a polynomial resource bound, or exact equality of all higher-dimensional records.

Proof. Take the fixed-$q_j = 7$ realizations of Theorem C.2 at the prescribed state, with their common endpoint S_*. At a limiting data set put an inactive sample at 0 and $d+1$ duplicate active samples at e_1, each with weight $1/(d+1)$. Initialize all added input weights at zero. The original loss is exactly the original loss; extra parameter Hessian directions are zero, and the duplicated observations are equal for every word. This is only an auxiliary limit, not the final construction.

For $\tau > 0$, replace the active samples by

$$x^{(0)} = e_1 + \tau \mathbf{1}_d, \qquad x^{(k)} = x^{(0)} + \tau e_k \quad (1 \le k \le d),$$

and initialize transverse weights by

$$W_{j,k} = \tau\big(w_j(0) + \mathbf{1}_{\{j=k-1\}}\big), \qquad k = 2, \dots, d,$$

leaving the original a, w_1, b, c unchanged. The augmented active square design has determinant $\pm\tau^d$, hence its weighted Gram determinant is $\tau^{2d}/(d+1)^{d+1} > 0$. A lower bound is this determinant divided by the dth power of the Gram trace. All input coordinates of each active sample are positive. For small enough τ, all active residuals and a_j remain positive, so every transverse weight derivative $-a_j \sum_{i \text{ active}} \lambda_i r_i x_{ik}$ is strictly negative. The initial weight rank is $\min(m, d)$: subtracting τ times the first column from the transverse columns leaves $\tau e_1, \dots, \tau e_{\min(m,d)-1}$; the remaining component of the positive first column is nonzero. For $m < d$ no width-m shallow model can have a larger weight rank.

Here are quantitative selection rules behind "small enough." The finite base paths have compact safety margin $s_0 > 0$. Choose disjoint closed windows about all events of each word. Off the respective window each comparison polynomial has absolute value at least $\gamma > 0$; on the window its time derivative has the required sign with magnitude at least $\nu > 0$. All minima are positive since there are finitely many paths and events. On a compact branch neighborhood choose a field Lipschitz constant L, data sensitivity C, and initial displacement bound $M\tau$. Gronwall gives

$$\sup_{t \le S_*} \|\Theta_\tau(t) - \Theta_0(t)\| \le E_\tau := (M + CS_*)e^{LS_*}\tau.$$

Let $L_g, L_F, L_{\dot F}, L_O$ bound, respectively, guard, pair comparison, comparison-derivative and observation perturbations on this neighborhood (data displacement is included in these constants). Choose $\tau > 0$ satisfying

$$L_g(E_\tau + \tau) < s_0/2, \quad L_F(E_\tau + \tau) < \gamma/2, \quad L_{\dot F}(E_\tau + \tau) < \nu/2, \quad 2L_O(E_\tau + \tau) < \varepsilon_{\mathrm{obs}}.$$

A first-exit argument gives common safety and existence, the sign/derivative bounds give exactly each prescribed word, and the observation bound gives the asserted tolerance. Hessian spectra are controlled by their branch-matrix continuity and Weyl's inequality, including the zero transverse modes at the auxiliary limit. The common endpoint is still S_*, not word-dependent clocks. Nonzero strict margins leave a positive perturbation allowance. Bounded forcing adds its integral times e^{LS_*} to E_τ and a direct $\|\nabla F_{ij}\|$ times forcing term to the comparison-derivative bound. Choosing its magnitude below all these positive margins preserves monotonicity and transverse orientation for continuous forcing.

The original slice moments at initialization remain exactly (c_*, D_*, E_*), because the transverse coordinates vanish on se_1. They are auxiliary slice moments, *not* the entire empirical state of the new full-rank data. This distinction prevents an approximate extension being mislabeled as an exact prescribed-state theorem. □

Remark Q.2 (GD version and limitations)**.** *For a supplied source common GD schedule with simple polygon crossings, consistent node-event orientations and a strict physical inter-event separation greater than one step, the same finite-update continuity argument gives a common-rate full-rank lift and independent update-error neighborhoods. The clock construction in Appendix C supplies this slack: its physical clock speed is at most* $10m$*, target clock gaps are at least* $\epsilon g_*/2$*, and* $\eta \le \epsilon g_*/(40m)$*, so physical inter-event gaps are at least* 2η*. The displayed theorem is stated for GF; the supplied safe/slack common schedule is an explicit hypothesis for this finite-step variant, and no polynomial bound in the word length is inferred. The active design becomes ill-conditioned as* $\tau \to 0$*. Consequently this theorem alone does not establish exact full-language realization at a prescribed nondegenerate higher-dimensional empirical state. The explicit Theorems O.1 and P.1, with design eigenvalue* > 0.14*, address that degeneracy objection for robust separation, not for every word.*

Q.1 New finite certificates and evidence boundaries

The planar and two-hidden-layer theorems combine the analytic tube, event-polynomial and output bounds above with finite outward arithmetic. The programs `certify_planar.py` and `certify_deep.py`, included under `supplement/code/`, implement the original nominal GD recurrence at fixed 160-bit dyadic precision, bound the continuum of prescribed data/initial/update errors, and write exact rational endpoints. These are not extrapolations from random perturbations. The integer implementation and its correspondence to the written inequalities remain trusted; neither theorem is an end-to-end Lean result.

The earlier certificate review reran both programs on new copies, without changing their code, configurations or published thresholds; the recorded certificate files match the inherited final certificates byte-for-byte. The archived `check_mechanism.py`, also included there, independently uses exact fractions for the first two planar updates, the shared Gram recurrence, direct kernel formula, noncommutator and loss of the old q_j invariant. A nonorthogonal negative control produces a different initial kernel. This finite exact test supports the displayed diagnostic, not the universal perturbation conclusions by itself.

The separately archived original-ReLU autograd check from that review contains 12 shallow and 16 deep strict states, including both sides of the training wall, with Jacobian/gradient/Hessian differences below $1.1 \cdot 10^{-14}$. Its equivariance check is a finite software diagnostic, not an exact trace proof. The inherited $5{,}469$ rational arithmetic tests likewise cannot prove the interpreter correct. Eight higher-dimensional GF illustrations have small design eigenvalues (about $3.1 \cdot 10^{-8}$ to $7.1 \cdot 10^{-8}$) and ordinary, not interval, ODE integration. They illustrate Theorem Q.1; they are not evidence of architecture-uniform conditioning.

None of these entries changes the $1{,}711$ frozen experimental records in Appendix M. Historical proof, nominal data, uncertainty bounds, stored-trajectory measurements, and kernel-checked Lean coverage remain distinct. No SGD, momentum, weight decay, arbitrary-depth full-history realization, or causal role of collisions is asserted. A machine `PASS` is tied to the exact parameters, schedules, uncertainty boxes and diagnostic norms checked here.

Relation to nearby results. Loss-preserving parameter transformations are studied by Zhao et al. (2023), while Grigsby et al. (2023) study additional symmetries of the full represented function. The common-training-signature isometry used here is proved directly in Lemma N.1; it need not preserve the function on every input. Local dimension and training geometry are studied by Bona-Pellissier et al. (2026). Grillo and Montúfar (2026) prove existence of open sets of parameters identifiable from the full function, up to neuron permutations and positive rescalings, when the input and hidden widths are at least two; this is not an almost-everywhere identifiability assertion for arbitrary parameters. Our exact pairs have different whole-domain functions and match empirical observations only, so there is no contradiction. The added contribution is the explicit nondegenerate planar and genuinely switching finite-budget separation, together with the correct asynchronous-wall observation contract, not a new symmetry, interval arithmetic method, or assertion that every network is unidentifiable.

R ADDITIONAL EXPERIMENTS: E4 AND E5

E4 adds 130 pairs and E5 adds 192. They are separate from the 1,711 original records in Appendix M. Each pair contains two networks, A and B, updated separately. Extra observations of saved networks do not count as new training runs.

R.1 SETTINGS AND PAIR CONSTRUCTION

Both models use $(0,0),(1,-1/2),(1,1/2),(2,0)$, zero targets and weights $(1,1/2,1/2,1)$. We use the half-sum loss in (1), without dividing by the sample count or total weight. The initial parameters are those of Examples 2 and 3, also specified in Appendices O and P. The orthogonal map constructs the initial centers only; it is not used to correct either trajectory. Full-batch GD evaluates the actual ReLU and updates all parameters together in float64.

Table 16: E4 settings. Data, initialization and total update budgets are coordinatewise bounds. Data include inputs, targets and weights. Total update error includes arithmetic effects as well as injected noise.

Model	Parameters	η	N	Data	Initial	Update
Planar	13	$1/200$	16	10^{-4}	10^{-4}	10^{-6}
Two hidden layers	23	$1/50000$	128	10^{-6}	10^{-6}	10^{-8}

Each model has one nominal pair and four groups of 16 perturbed pairs. The groups turn shared data perturbations and independent update noise on or off; all use independent A/B initialization errors. Fixed PCG64 streams generate uniform-grid perturbations, with amplitudes 99% of the data/initialization budgets and 50% of the update budget. Integer draws and binary arrays are saved. Groups use different initialization streams. These are local perturbation checks, not samples from a standard random-network initialization law.

Separate checking code recomputes $\theta_{k+1}-\theta_k+\eta\nabla L_d(\theta_k)$ from the saved binary states using exact rational arithmetic. All 130 pairs satisfy their budgets. The test rectangles are fixed for both E4 and E5:

$$J_P=[0.76,0.82]\times[-0.05,0.05],\qquad J_D=[0.61,0.64]\times[-0.01,0.01].$$

The two exact nominal functions agree on their rectangle *at initialization*. Their final and training-increment differences are separate quantities. For each saved pair, a common piecewise-affine subdivision of $f_A^0, f_B^0, f_A^N, f_B^N$ gives the rectangle extrema. The 201×201 heatmaps illustrate these functions; pixel minima are not used as uniform bounds.

R.2 TRAJECTORIES AND DIAGNOSTIC DIFFERENCES

Figures 9 and 10 show the nominal pairs. Labels follow the initial breakpoint order. Root checks cover each saved parameter segment, not just its endpoints. Planar collision polynomials have degree at most two; on the verified two-hidden-layer probe segment they have degree at most four. The histories concern $y=0$, not the full planar region adjacency structure.

For predictions, the empirical NTK and the sorted Hessian spectrum, we report the largest absolute entry difference. For loss and the two norms, we report the absolute scalar difference. The NTK is $JJ^\top$, without sample-weight or sample-count normalization; only the loss includes the training weights. Saved readouts include nodes, quarter-step fractions, a common 257-point grid and observations near events. A fractional step is a linear interpolation of saved parameters, not an extra GD update. Plotted maxima are sampled values, not continuous-time upper bounds.

In the two-hidden-layer nominal pair, roundoff separates the two lower-gate crossing times. The nodewise NTK difference is about 3.55×10^{-14}, but between the crossings it reaches about 0.04378. We retain this peak and do not align the trajectories in time. Node/window-exterior differences, synchronous sampled maxima and time-average estimates are distinct, as in Appendix P. We do not assign a classical Hessian at the boundary.

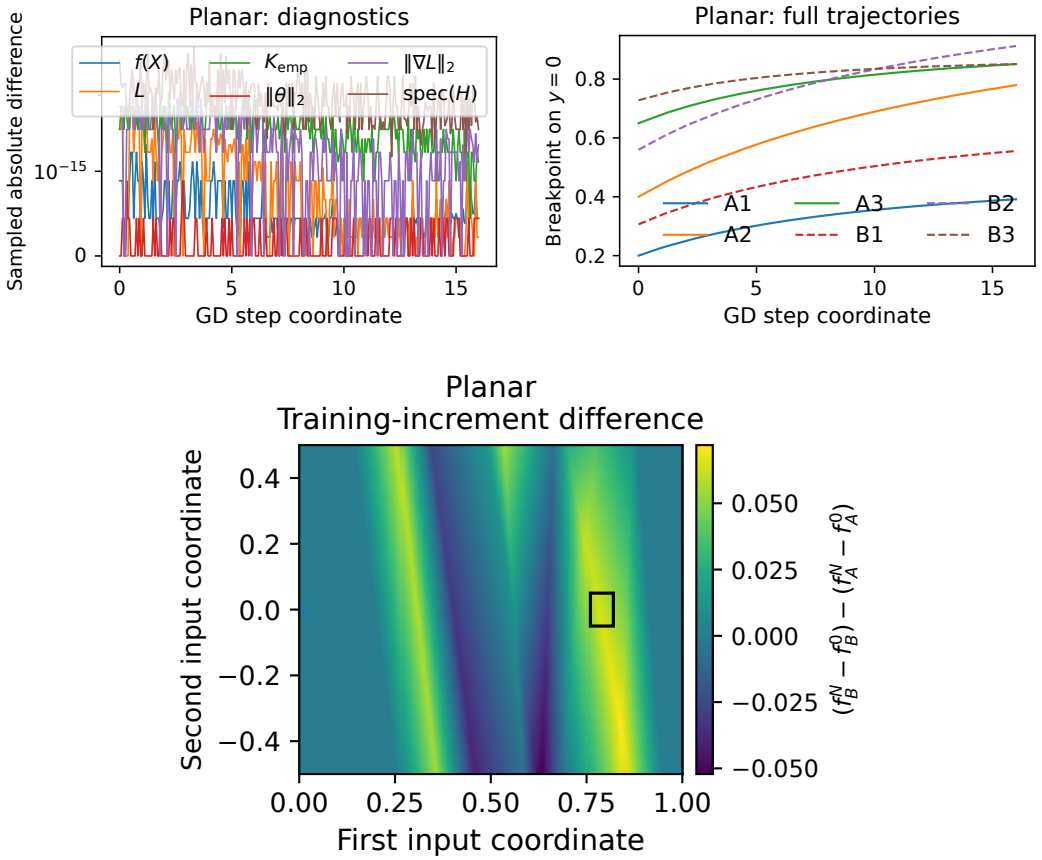


Figure 9: Planar nominal pair: sampled diagnostic differences, tracked breakpoint positions, and $(f_B^N - f_B^0) - (f_A^N - f_A^0)$ on $\Omega = [0, 1] \times [-1/2, 1/2]$. Solid breakpoint curves are A and dashed curves are B. Only B has a $(2, 3)$ collision. The rectangle is J_P; circles mark training inputs in the displayed domain. The heatmap does not certify a minimum.

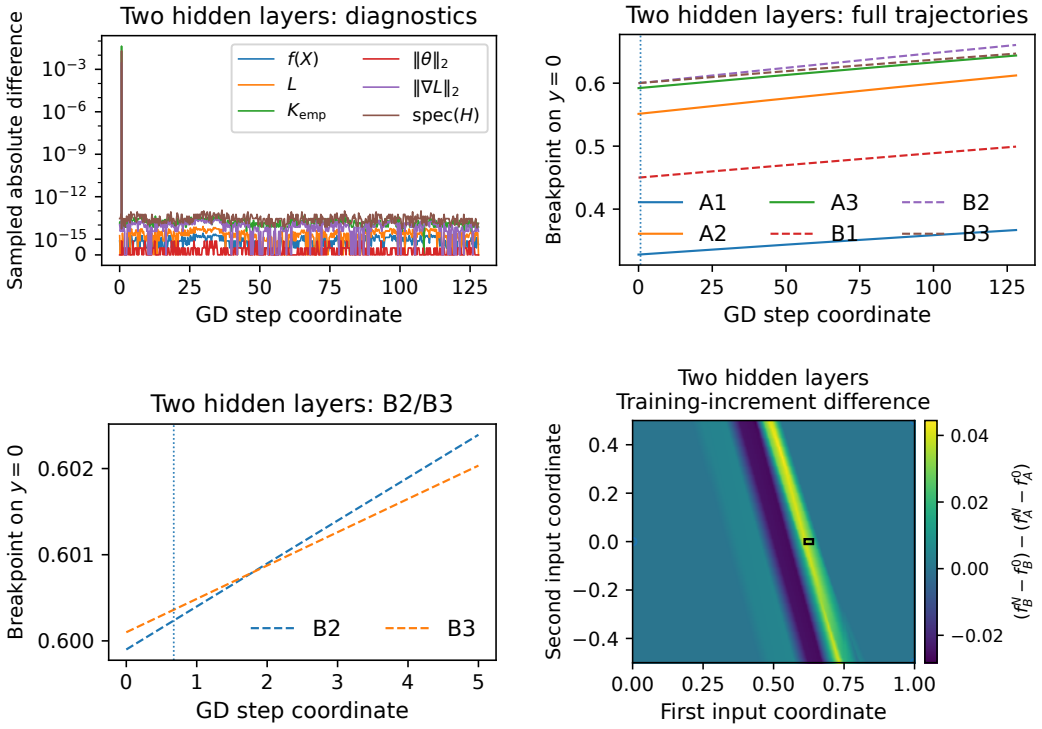


Figure 10: Two-hidden-layer nominal pair: sampled diagnostic differences, tracked breakpoint positions, an early-time view of B2/B3, and the training-increment difference on Ω. Dotted lines mark the lower-gate changes in A and B, which are too close to separate visually. B's later $(2, 3)$ collision is a different event; A has no tracked collision. The heatmap rectangle is J_D. The diagnostic peak near the training boundary is retained.

R.3 Added samples and group comparisons

The mixed block consists of the three planar hidden units or the first three upper units in the two-hidden-layer network. A common pattern means that, at each training input, the block is all active or all inactive, with the same choice in A and B. Other layers need not keep fixed masks.

For each model and each of 16 seeds, we select 28 pairs of added points. Common candidate numbers $u, v \in (0, 1)$ give

$$(1.2 + 0.3u, -0.1 + 0.2v) \quad \text{and} \quad (0.3 + 0.4u, -0.1 + 0.2v).$$

We accept a pair if the first point activates the whole mixed block in both networks, the second gives mixed signs in both, and every hidden preactivation has magnitude at least 10^{-4}. We reject duplicates. Selection uses only initialization, not subsequent training. The first 4, 12 and 28 accepted points give 8, 16 and 32 total samples.

For each enlarged dataset, the original four weights are multiplied by $3/4$. Added points each have weight $3/[4(n-4)]$. Thus the original and added weights sum to 2.25 and 0.75, and the total remains 3. Targets are zero. The groups share the sample count, initial centers and schedule, but not the first-coordinate distribution. This is a matched comparison, not an isolated causal test. E5 adds no initialization or update noise; roundoff remains. The four-sample baseline reuses the two E4 nominal pairs without new runs. E5's enlarged data are outside the E4 perturbation boxes, so the E4 gap thresholds are not E5 acceptance criteria.

Table 17: E5: median final rectangle minimum across 16 data seeds per group. Unresolved counts incomplete sign classification near training roots, not failed training, rectangle evaluation or slice-history verification. P/S means Preserve/Split.

Model	Samples	Preserve	Split	Unresolved (P/S)
Planar	8	0.051329	0.030266	0/0
Planar	16	0.052104	0.030953	0/0
Planar	32	0.052325	0.030836	0/0
Two hidden layers	8	0.023259	0.004489	0/15
Two hidden layers	16	0.023248	0.004479	0/16
Two hidden layers	32	0.023226	0.004412	0/16

All 192 pairs have positive final and training-increment rectangle gaps. A has no tracked collision and B has one $(2, 3)$ collision. All 96 preserve pairs retain the common block pattern, and they have larger final gaps than their matched split pairs. Split pairs are already distinguishable at initialization: the maximum added-input prediction difference ranges from 0.02875 to 0.05585 for the planar model and 0.02014 to 0.03719 for the two-hidden-layer model. They are not additional identical-record examples. Across sample counts, datasets use nested prefixes of the same 16 pools; across groups, they use paired locations. We report descriptive results, not an independence-based success probability.

Unresolved sign intervals. The analysis distinguishes a sign split on an open interval, a zero contact, and failure of strict common signs. A contact need not be a crossing. All split pairs fail common signs at initialization. In 47 two-hidden-layer split pairs, 385 intervals around training roots retain an unresolved sign classification. Their maximum optimization-time width is below 2.0×10^{-30}. Root existence, uniqueness within each interval and completeness were checked separately. We do not replace these intervals by a fixed sign or discard them. A short interval can still contain a diagnostic jump. Rectangle extrema and tracked slice histories are verified independently for all 192 pairs.

R.4 Extra observations without retraining

The `READOUT_ONLY` control evaluates added points on the unchanged E4 nominal trajectories. It changes neither the loss nor the parameters. Each model/group/seed combination supplies a 28-point set; its first 4, 12 and 28 points give the three readouts. These 64 sets and 192 prefix evaluations are not trained pairs. Split locations already distinguish predictions at initialization, so extra observations can reveal information without retraining. Parameter Jacobians are auxiliary readouts, not components of (2); equal NTK matrices need not have equal Jacobian factors.

Table 18: READOUT_ONLY: largest added-point prediction differences across 48 prefix evaluations per row. Initial and Final use $f_B - f_A$; Increment uses $(f_B^N - f_B^0) - (f_A^N - f_A^0)$. These are maximum absolute entry differences, not the signed rectangle minima in Table 17.

Model	Added locations	Initial max.	Final max.	Increment max.
Planar	Preserve	1.78×10^{-15}	8.88×10^{-16}	2.22×10^{-15}
Planar	Split	0.055851	0.038044	0.055851
Two hidden layers	Preserve	8.88×10^{-16}	1.33×10^{-15}	2.00×10^{-15}
Two hidden layers	Split	0.037186	0.031478	0.026218

Verification scope. Separate checking code recomputed saved update residuals and rectangle extrema for all 322 pairs. It checked E5 event intervals; E4 event checks came from the preceding review. Dense diagnostics were independently recomputed for 14 selected pairs and endpoints for all 322. The other 308 dense time series and time-average quadrature arrays were not all independently recomputed. Read-only checks cover all 64 sets' endpoints and four complete provided time series. These checks concern saved finite objects. They are not a new universal theorem, a continuous-time error certificate, or an end-to-end Lean verification. This manuscript revision uses the archived data without retraining or changing the test rectangles or seeds.