FROM CLASS MEANS TO MARGINS: PREDICTING HIDDEN-UNIT PLACEMENT IN WIDTH-ONE NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

At small output scale, logistic loss favors separation of class means; at large scale, it favors the worst-case margin. We study a width-one network with activation $f_a(t) = t + a \sin t$ on symmetric inner-versus-outer intervals, the smallest setting in which these preferences require different hidden features. An independently rechecked certificate locates the placement switch, and a scaling reduction predicts its activation dependence with a first-order coefficient confirmed at finite parameter values. Registered prospective tests predict training crossings from conditional thresholds fixed before training. A first-order tracking law, derived post hoc, predicts each run's lag from the branch geometry and its measured ratio of output-scale growth to hidden relaxation. At an activation value with no prior crossing data, its registered test gives median per-run observed/predicted ratios of 1.07 under Adam and 1.06 under SGD; individual predictions are accurate under SGD, where every run is within 10%, but not under Adam, where 22% are. A registered learning-rate test finds that the lag does not vanish as the learning rate decreases. This quantitative predictive link is established for width-one sine networks that reach a branch and grow slowly relative to relaxation; outside this setting, registered predictions fail or remain unresolved, and agreement conditioned on the eventual branch is explanatory.

1 INTRODUCTION

At small output scale, logistic loss rewards a feature that separates the class means. At large scale, it rewards separation of the hardest points. We study what happens when these preferences require different features: a width-one network $N(x) = w_2 f_a(w_1 x + b_1) + b_2$, with $f_a(t) = t + a \sin t$, trained on $y = \operatorname{sign}(|x| - 1)$. On symmetric inner-versus-outer windows, the mean-separating feature has a steep input weight whose linear ramp prevents correct placement. Increasing output scale changes the preferred feature. Adam and full-batch SGD follow this change with a measurable lag.

An independent ball-arithmetic checker, sharing no code with the searches, re-checks the placement certificates. Predictions are fixed in timestamped registrations before the scored data are collected, and a ledger recomputes every reported number from committed artifacts. In width-one sine networks, the fixed-scale minimizer predicts placement up to a computable tracking lag.

The fixed-scale minimizer traces a regularization path; tracking it is an established idea. We compute its placement switch and tracking lag in a non-monotone network. At $a = 1.65$, with no prior crossing data, the test confirms typical lag under both optimizers and individual ordering under SGD; Adam's ordering criterion fails. The calculation uses loss derivatives and an optimizer metric. The original Adam metric came from measured runs; the new test freezes its measurement rule before training.

The minimal setting permits an exact margin decomposition and global conditional-loss certificates unavailable in our larger-network experiments. Non-decreasing activations cannot solve the task, giving an exact representational boundary distinct from the training threshold. The selection criterion applies beyond width one under explicit hypotheses. At width two it predicts the conditional landscape, but training need not follow that landscape.

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.

- The margin identity separates placement from bias alignment.
- The selection criterion and sine-specific bracket establish the feature change; certificates locate the width-one switch.
- The scaling reduction predicts geometry dependence without fitting; independently checked thresholds confirm its first-order coefficient.
- Prospective tests predict width-one crossings and typical Adam and SGD lag, while rejecting Adam's individual ordering.
- Learning-rate and transfer tests delimit tracking: a branch must be reached, with output growth slow relative to relaxation.

2 SETTING AND THE MARGIN IDENTITY

The negative window is $I = [-0.8, 0.8]$ and the positive windows are $O = [-2.0, -1.2] \cup [1.2, 2.0]$. Correctness requires strict signs throughout. The base protocol uses full-batch Adam, mean logistic loss, 200 points per class, learning rate 10^{-2} and initialization $U(-1, 1)^4$. Headline sweeps are float32; conditional analyses are float64. We use an 800-point objective (a fixed quadrature of the input distribution, which we call the population objective), distinct from continuous-window correctness (Appendix B).

Proposition 1 (Monotone impossibility)**.** *No width-one network with a monotone activation classifies these windows correctly.*

Proof. Its output is monotone, whereas correctness requires positive output on both sides of a negative interval. □

The sine activation is non-decreasing exactly when $|a| \le 1$. We study $a > 1$, with $\varepsilon = a - 1$. Orient the hidden output as $u = \operatorname{sign}(w_2) f_a(w_1 x + b_1)$ and write $s = |w_2| > 0$. Define

$$U = \max_I u, \quad V = \min_O u, \quad G = V - U, \quad G^\star(a) = \sup_{w_1, b_1} G. \tag{1}$$

Oddness makes $(w_1, b_1, w_2) \mapsto (-w_1, -b_1, -w_2)$ preserve the network. Let $\widehat{G}$ be the checker's lower bound at the retained gap witness and $R = s\widehat{G}/2$; $G^\star$ is enclosed separately. The exact margin identity is

$$\gamma = \min\{sV + b_2, -sU - b_2\} = \frac{sG}{2} - \left| b_2 + \frac{s(U+V)}{2} \right| = R\rho - |c|, \tag{2}$$

where $\rho = G/\widehat{G}$ and $c = b_2 + s(U+V)/2$. Correctness is exactly placement, $G > 0$, plus bias alignment, $b_2 \in (-sV, -sU)$. Output scale is absent from the placement condition. Joint positive rescaling of (w_2, b_2) preserves decisions, so correct networks exist at arbitrarily small scale with small margin. For margin $m > 0$, subtraction gives $s \ge 2m/G \ge 2m/G^\star$ (Appendix C).

Of the 2,400 headline runs, 430 solve, 1,664 have incorrect placement and 306 have incorrect bias alignment. Placement accounts for 84.5% of unsuccessful runs (Figure 1). Scale affects how placement is learned.

3 THE FIXED-SCALE DICHOTOMY

Fix s and minimize loss over the hidden parameters and output bias. Profiling the unique optimal bias leaves a two-dimensional problem. On balanced data its small-scale expansion is

$$\mathcal{L}^*(\theta; s) = \log 2 - \frac{s}{4}\Delta\mu(\theta) + \frac{s^2}{8}\operatorname{Var}(u_\theta) + O(s^4), \quad \Delta\mu = \mathbb{E}_O u_\theta - \mathbb{E}_I u_\theta. \tag{3}$$

The mean gap controls the leading preference; variance selects among ties. At large scale, exponential loss bounds favor the worst-case data gap G_n, the sample version of G.

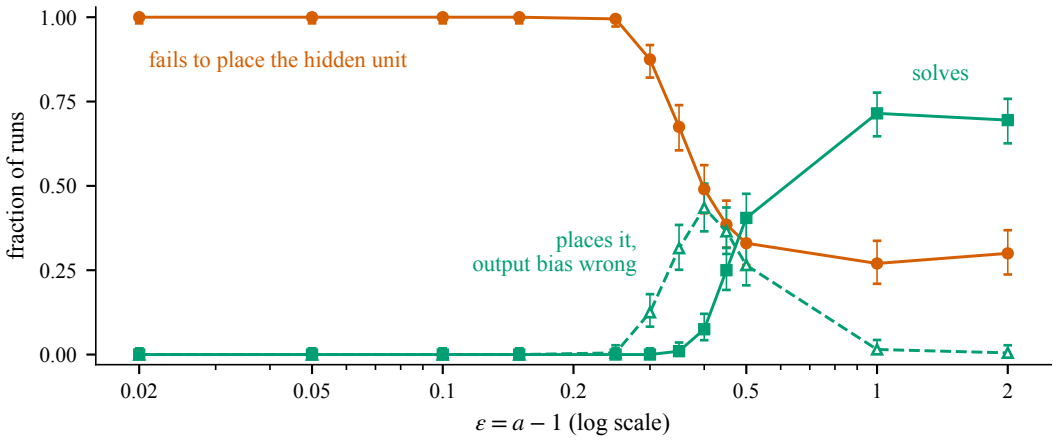


Figure 1: Most unsuccessful runs have incorrect hidden-unit placement. The single-draw float32 sweep contains 2,400 width-1 runs, 200 at each of 12 activation values, trained with Adam at learning rate 0.01 for 2,000 steps: 430 solve, 1,664 have incorrect placement and 306 have incorrect bias alignment. Bars show Clopper–Pearson 95% intervals. These are the runs used in the continuous-metric check.

Theorem 1 (Small- and large-scale selection)**.** *Assume the class-mean gap attains its maximum, and that conditional minimizers remain, as $s \to 0$, in a common compact set on which u_θ is continuous and bounded. Every small-scale limit point maximizes the mean gap and minimizes variance among those maximizers. If all selected points have strictly negative G_n, and a positive maximal data gap is attained, conditional minimizers are unplaced for sufficiently small s and placed for sufficiently large s. If minimizers also stay compact as $s \to \infty$, their limit points maximize G_n.*

The criterion is sufficient; a converse and a unique crossing are not established. For the sine family, attainment and large-scale compactness are proved. Small-scale compactness remains assumed, since fixed-s localization gives no common compact set as $s \to 0$. The proof uses an explicit remainder bound (Lemma 4) and needs no growth condition.

On symmetric windows the linear ramp contributes nothing to $\Delta\mu$, yet a steep input weight maximizes the mean gap. Endpoint pairs impose $G, G_n \le 2a - 2.8|w_1|$, so that feature is unplaced. Convexity and this bound give an explicit sine-specific switch bracket without the small-scale compactness assumption: at $a = 1.30$, an unplaced candidate beats every placed one for $s < 0.342$, and placed solutions beat every data-unplaced one for $s > 163.8$ (Appendix A). Figure 2 shows how the conditional minimizer moves between them. Placement changes along a continuous parameter branch. The finite quadrature selects an aliased class-mean maximizer at extremely large input weight. It too is unplaced, preserving the width-one conclusion. The width-two small-scale selection statement is made on the continuous windows (Appendix A.2).

4 Certificates and the Scaling Law

Write R_{glob} and R_{solve} for the population conditional minimizer's placement and sign-correctness thresholds. On a run's own sample, R_{own} denotes the global placement threshold and $T^{\pm}_{\mathrm{own}}$ the thresholds of its mirror branches. The thresholds depend on the geometry and objective. Periodicity and an analytic input-weight bound make the fixed-scale search compact; its audit found no lower-loss candidate among 16,468 searches.

The finite placement brackets, gap enclosures and limit-switch bracket are independently verified in ball arithmetic, as are localization, the cubic-gap maximum K, the first-order Krawczyk boxes and the positive-definite and ring checks. Outer exclusion and the finite and limiting solve brackets remain search-certified only. These checks verify the limiting interval $[0.19738, 0.19920]$; Table 1 gives the sharper value's status (Appendix E.1).

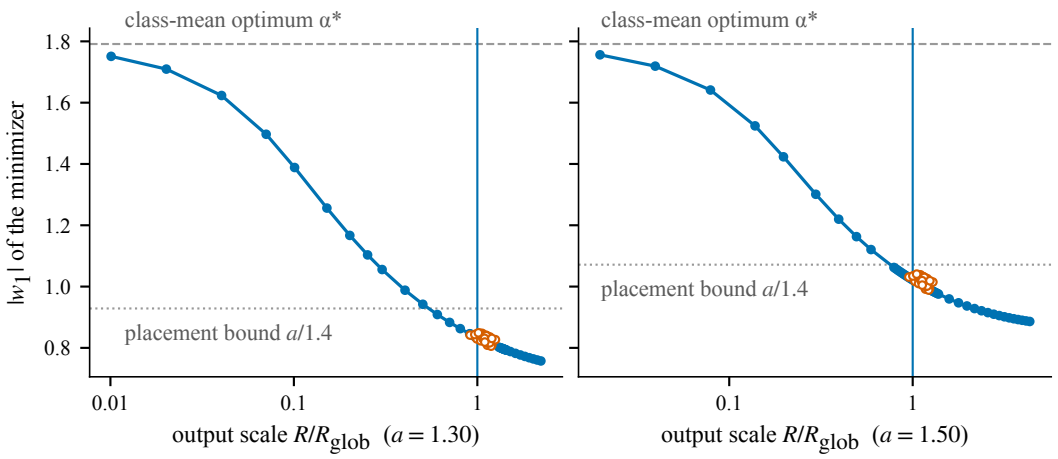


Figure 2: As output scale grows, the conditional minimizer leaves the class-mean optimum and drops below the necessary placement bound before its gap changes sign. Width 1, $a = 1.30, 1.50$, 800-point population objective. The minimizer path is a validated search, not a certificate; R_{glob} is the certified bracket midpoint. Free training gives 38 crossings per value at 32,000 steps, checked every step. The path is a deterministic minimization, so no interval is drawn; crossings are individual runs. The mean-gap optimum in the plotted range is $\alpha^* = 1.7913$; smallest-scale input weights are 1.751 and 1.756. The bounds $a/1.4 = 0.929, 1.071$ are crossed at $R/R_{\text{glob}} \approx 0.54, 0.76$; they are necessary, not sufficient. All training crossings satisfy the bound. At R_{glob} the output weight lies in the certified brackets $|w_2| \in (4.9500, 4.9625]$ at $a = 1.30$ and $(2.5250, 2.5375]$ at $a = 1.50$.

Table 1: Status of the main claims. Proofs apply under their stated hypotheses.

Claim	Status
Margin identity; analytic switch bracket; small-scale selection (Theorem 1)	Proved.
Finite-activation placement thresholds	Computer-assisted, independently re-checked.
Sharp limiting threshold	Computer-assisted, not fully independently re-checked.
Branch scale law, Equation (5)	Proved for the continued branch; tied to the global threshold by certificates at finite activation values.
Lag law	First-order approximation, derived post hoc.
The $a = 1.65$ test	Registered; uses each run's measured output-scale growth; accurate per run under SGD, in the median under Adam.

In dip coordinates $w_1 = \sqrt{\varepsilon}p$, $b_1 = \pi + \sqrt{\varepsilon}q$,

$$\frac{f_{1+\varepsilon}(\pi + \sqrt{\varepsilon}\sigma) - \pi}{\varepsilon^{3/2}} = h(\sigma) + \varepsilon r(\sigma) + O(\varepsilon^2), \quad h = -\sigma + \sigma^3/6, \quad r = \sigma^3/6 - \sigma^5/120. \tag{4}$$

Proposition 2 (Scaling reduction)**.** *H1 localizes the maximal-gap placement, up to symmetries, in $\mathcal{C} = \{1 \le |p| \le 2.2,\ |q| \le 2\}$ and is proved for $0 < \varepsilon \le 0.029$. Then $G^\star(1+\varepsilon) = K\varepsilon^{3/2}(1 + O(\varepsilon))$, with $K = \sup G_0$ for h. On bounded rescaled sets, the profiled loss depends on output scale only through $sG^\star/2$, up to $O(sG^\star\varepsilon)$. For the base quadrature, localization and global uniqueness of the limiting minimizing branch, up to reflection, are certified through the switch. H2′ is its nondegenerate, transversal gap crossing. These give local persistence for sufficiently small positive ε; global finite-a preference uses its separate certificate.*

H1 concerns the gap maximizer; identifying its tied maximizing corner remains certified. Replacing $G^\star$ by $\widehat{G}$ adds the enclosed geometry error. Along the continued branch, the switch scale is

$$s_{\text{br}} = A^*\varepsilon^{-3/2}(1 + c_s\varepsilon + O(\varepsilon^2)), \quad A^* \in [0.68125, 0.6875], \quad c_s = 0.66215. \tag{5}$$

At every certified activation value, the global threshold lies within 3% of this branch switch, and in the certified limit the two coincide. The limiting geometric constant predicts measured constants

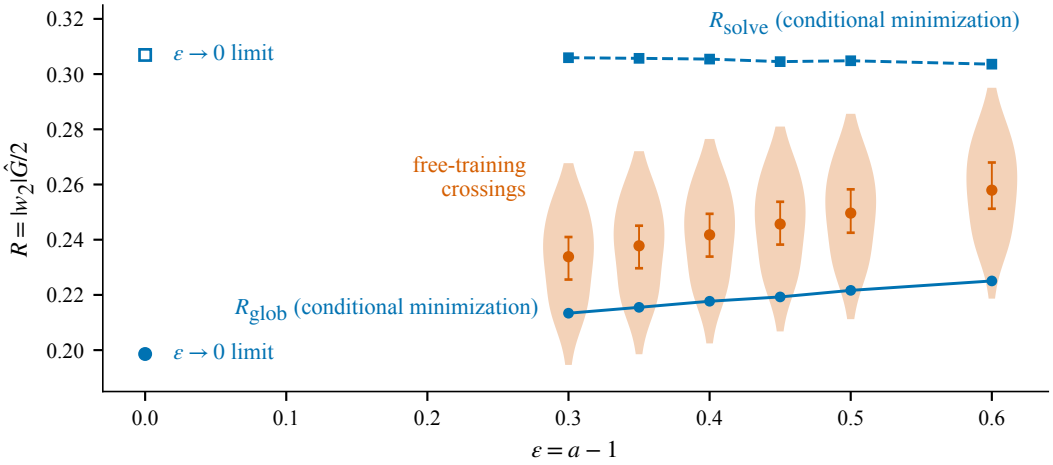


Figure 3: At each tested activation value, free training acquires placement above R_{glob} and below R_{solve}. There are 38 crossing runs at each of six values $a = 1.30$–1.60, at budget 32,000 with every-step detection. Violins show crossings; points are medians with bootstrap 95% intervals. Certified brackets, plotted at their midpoints, have widths at most 1.4×10^{-3} for either threshold. The sharp limiting placement value $[0.1985926, 0.1985927]$ and rounded limiting solve value $[0.3067, 0.3071]$ use the global cubic-gap normalization of Lemma 2; their verification status is given in Table 1 and the text. The follow-up passes its registered rule (Appendix F.2); the original test exceeds its pre-set width.

within 0.808% across five new geometries spanning an 8.1-fold range, with nothing fitted. The original cross-family solve-threshold comparison misses its criterion; the follow-up restricted to the polynomial domain agrees (Appendix Q).

In R units, $c_1 = 0.28523$ is confirmed by certified thresholds at $\varepsilon = 0.08$–0.12, with all ten endpoint certificates independently re-checked. The feasible set $[0.26531, 0.30687]$ contains the prediction and excludes the competing values. The original small-ε test remains inconclusive. The branch law leaves output growth and tracking lag undetermined (Figure 3).

5 Training tracks the minimizer

Prospective thresholds. Conditional thresholds fixed before training improve crossing prediction on held-out geometries and fresh samples. A geometry predictor uses a lag factor fitted on the base window; the own-sample predictor also improves over the population threshold without calibration (Figure 4). Earlier geometry predictions missed their criteria, and a later offset-range prediction misses one setting. A stronger pooled-lag baseline matches the calibrated predictor, while threshold-free baselines do 4 to 5 times worse; that comparison was made after observing the data (Appendix I.1).

The SGD own-sample result is a partial pass: among the 30 of 40 crossers at each of $a = 1.30, 1.50$, own thresholds improve prediction and meet the rank-correlation bound; including non-crossers preserves the improvement but rejects the correlation criterion. Fresh samples at $a = 1.45, 1.60$ confirm both own-threshold improvement and the earlier fitted timescale relationship, with slight extrapolation at $a = 1.60$ (Appendix L).

5.1 The tracking lag

Let $\theta^*(s)$ be the occupied stationary branch in trained coordinates (w_1, b_1, b_2), with positive-definite Hessian H. Its motion displaces the hidden parameters from equilibrium. The first-order slow-manifold correction of geometric singular perturbation theory (Fenichel, 1979) gives

$$\delta_{t+1} - \delta_t = -\eta P H \delta_t - \theta^{*\prime} \dot{s}, \qquad \delta \simeq -(\eta P H)^{-1} \theta^{*\prime} \dot{s}. \tag{6}$$

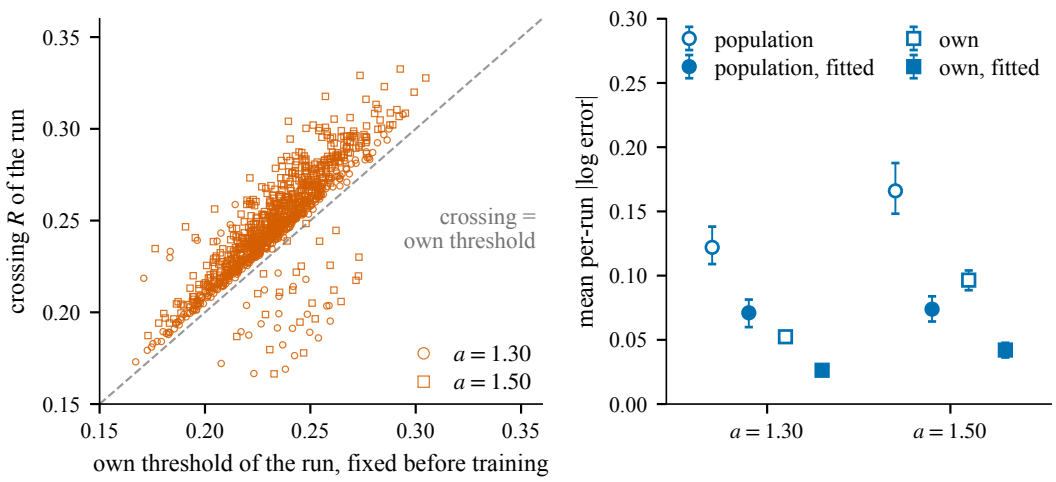


Figure 4: A run's own threshold, fixed before training, improves crossing prediction over the population threshold, with or without calibration. Sixteen new settings have 60 runs each; 460 of 480 cross at each activation value, detected at 50-step checks. In (a), frozen own-threshold brackets have width at most 8.4×10^{-4} and are not drawn; per-setting Spearman correlations are 0.76–0.86 at $a = 1.30$ and 0.57–0.82 at $a = 1.50$. In (b), predictors are R_{glob}, fitted $\lambda(a)R_{\text{glob}}$, R_{own} and fitted $\rho_{\text{lag}}(a)R_{\text{own}}$, with $\rho_{\text{lag}} = 1.0311$ and 1.0660 at the respective activation values. Bars are window-bootstrap 95% intervals from 10,000 resamples, seed 0. All registered criteria pass under both detection definitions; the claim that fitted population prediction beats the unfitted own threshold at $a = 1.50$ does not survive interpolation: difference $-0.003\,[-0.013, 0.008]$ (Appendix I.2).

Here $P = I$ for SGD; Adam's preconditioner is treated as stationary on the relaxation timescale. Linearizing the gap at the branch switch s^* yields

$$r = \frac{s_c - s^*}{s^*} \simeq \kappa(a)\chi, \quad \chi = \frac{\dot{s}/s^*}{\eta\lambda_{\min}}, \quad \kappa = \lambda_{\min}\frac{\nabla G^\top (PH)^{-1}\theta^{*\prime}}{\nabla G^\top \theta^{*\prime}}, \tag{7}$$

where $\lambda_{\min} = \lambda_{\min}(P^{1/2}HP^{1/2})$. The continuous-time theory supplies no tracking guarantee for discrete Adam or SGD. Tests using $\dot{s}/s_c$ differ in normalization by $1 + r$.

At the tracked branch's switch, the closed form predicts all 36 arm medians within 10%; integrating the linearized dynamics along measured output-scale paths brings every defined arm median within 5% (Figure 5). Using the global own-sample threshold instead gives only 31 of 36 closed-form arm medians within 10%. The observed crossing branch explains the difference retrospectively.

The Hessian, branch tangent, gap gradient and switch come from the loss landscape. In the original calculation, the winding and Adam preconditioner shape come from existing runs at crossing; each run also supplies its growth rate and metric for χ. Appendix V.5 lists the sources. Increasing sample size shrinks only the finite-sample threshold shift; the lag behind each sample's own threshold stays unchanged.

Registered test at an unseen activation value. At $a = 1.65$, all rules and per-seed thresholds were frozen before training 80 fresh seeds per optimizer. The switch is validated, not certified. The winding rule was defined from crossings at other activation values. The last upward passage of half the frozen threshold before that threshold was the pre-specified measurement point. Branch and preconditioner predictions use information available there; the trajectory-integrated calculation also uses the measured output-scale path. Predictions were committed before every-step crossing detection.

For both versions and both optimizers, the median per-run observed/predicted ratio is within about 7% of 1 (Table 2). Under SGD every predicted run is within 10% and the rank correlation is 0.995. Under Adam, 22% of runs are within 10%, the registered ordering criterion fails (rank correlation 0.25), and the ratio of median lags, 0.094 observed to 0.080 predicted, is larger than the median ratio, since the two statistics differ. Appendix V.7 gives exclusions and all criteria.

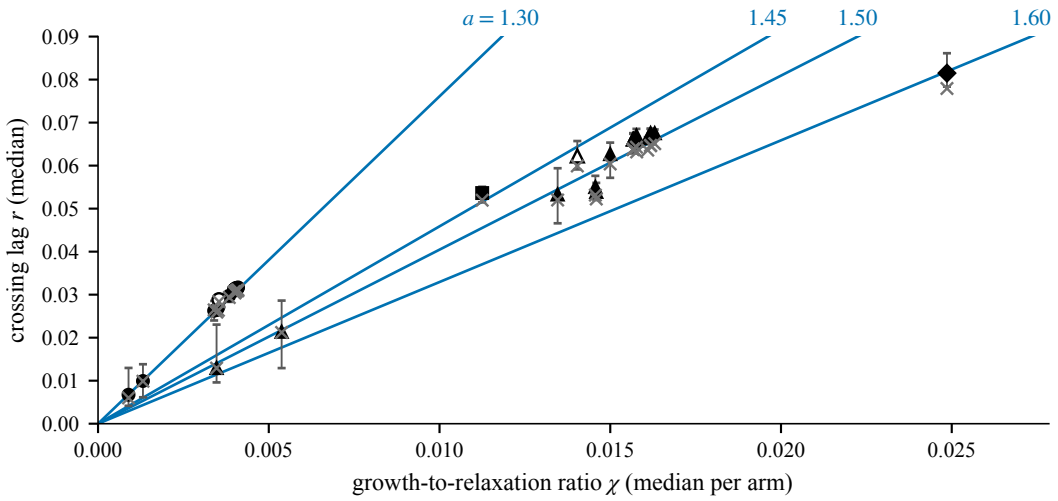


Figure 5: Lag grows with output growth relative to branch relaxation. All 36 width-one free-training arms are shown, measured from each run's tracked-branch switch; filled markers are Adam, open markers SGD. Bars are run-level bootstrap 95% intervals of arm medians (10,000 resamples, seed 20260926). Lines are predictions with no fitted parameter: $\kappa(1.30) = 7.61$, $\kappa(1.45) = 4.59$, $\kappa(1.50) = 4.05$, $\kappa(1.60) = 3.29$. Original global-own-threshold slopes are 0.89, 0.83, 0.91 and 0.82, with 36/36 arms inside their original tolerance. Grey crosses integrate linear response from $0.7s^*$: observed/predicted arm medians span 1.00–1.05 among 1,708 of 1,750 runs with a prediction. Both analyses were derived after the fitted relationship was known; predictions were committed before comparison.

Post hoc, the Adam failure traces to when the preconditioner was measured: freezing it at the observed crossing instead of at half the switch scale raises the per-run Spearman to 0.99, with every run within 10%, showing that the rule-point preconditioner was stale by the crossing. A pre-crossing rule chosen after seeing these runs, taking the preconditioner at the first step the output scale reaches the occupied branch's switch, gives 0.95 (91% within 10%); we name it as the rule for a future registered test.

Post hoc, with nothing fitted to these runs, a constant lag carried over from $a = 1.30$–1.60 underpredicts by a factor of 2.2–2.7. Under SGD the trajectory-integrated prediction is the most accurate (median absolute error 0.005, every run within 10%, rank correlation 0.99); under Adam the earlier fitted relationship does better than the registered prediction (0.010 against 0.023; 45% against 22% within 10%), consistent with the stale rule-point preconditioner.

Learning rate and interventions. The registered free-training test at learning rates 0.01, 0.005 and 0.0025 finds unchanged lag within tolerance: reducing the learning rate scales growth and relaxation together. This supports a nonvanishing gradient-flow lag over the tested rates. Adam's preconditioner matters; its forced-ramp test violates the stationary-metric assumption, while SGD's forced-ramp results are mixed (Appendix W). Held-scale tests predict the fraction placed. Individual replays restart from a saved state at fixed output scale: the starting branch's threshold predicts their outcomes more closely than the run's global threshold. Reducing scale removes placement in all 37 eligible runs, while retention above it, 33 of 37, falls short of the registered criterion. Supplying large output scale early gives the opposite of the registered prediction: gradients collapse and placement stalls, with 9 of 40 placed versus 37 of 40 controls. Longer training partly rescues these runs, and the cold-start variant is mixed (Appendices G and J).

6 Where the account applies

The predictive link is established for width-one sine networks on the interval geometries tested under Adam and SGD, with prospective own-sample tests at $a = 1.30, 1.45, 1.50, 1.60$ and the lag test at

Table 2: Registered lag test at $a = 1.65$, with every-step detection. Ratios are medians of per-run observed/predicted lag, not ratios of arm medians. Required ranges: integrated $[0.90, 1.10]$, closed form $[0.80, 1.20]$; required rank correlation ≥ 0.5.

	Adam	SGD
Crossers / initialized runs	76/80	64/80
Runs receiving predictions	74	63
Observed median lag	0.094	0.094
Predicted median lag, integrated	0.080	0.090
Predicted median lag, closed form	0.089	0.088
Median per-run ratio, integrated	1.065	1.056
Median per-run ratio, closed form	1.045	1.024
Runs within 10% (integrated)	22%	100%
Median absolute lag error	0.023	0.005
Rank correlation (registered; required ≥ 0.5)	0.25 (fail)	0.995 (pass)

$a = 1.65$. A run must reach a stationary branch and grow slowly relative to its relaxation. In a registered test with forced SGD ramps at five preset rates ($\chi = 0.03$ to 0.4, $a = 1.30$ and 1.50), both criteria were unresolved because most fast-ramp runs outran the tracked branch. In the three cells with enough crossings, the lag stayed within 25% of $\kappa(a)\chi$ where the predicted lag $\kappa\chi$ was about 0.1, and exceeded it by a factor of 1.7 at $\kappa\chi \approx 0.23$. These three cells suggest, post hoc, a condition on predicted lag, $\kappa\chi \lesssim 0.1$, rather than on χ alone. Across the 36 free-training arms, the predicted lag has arm medians from 0.007 to 0.082, with 95% of runs between 0.010 and 0.081, so every arm lies below that level.

At width two, cancelling the linear ramp makes the symmetric-window conditional minimizer placed at every scale searched; the small-scale selection statement applies on continuous windows. Asymmetric windows have a validated conditional threshold. Their full-speed training comparison and the symmetric no-gating prediction fail; the timescale-matched follow-ups remain unresolved, and the slowed own-width-one comparison misses its registered advantage. Agreement on a tracked branch outside width-one sine is post hoc, and branch selection could not be predicted in advance (Appendices N and X).

GELU, SiLU and Mish have validated conditional switches, but their primary training tests remain unresolved and their larger secondary tests reject the predictions. SiLU and Mish grow too fast for the observed tracking regime. A single unit on a noisy band task inherits the population threshold analytically; its $\mathbb{R}^2$ lag lies outside the predicted range, and the primary $\mathbb{R}^4$ test remains unresolved. The retrospective own-branch analysis is in Appendices Y.1 and Y.2.

In an exploratory extension to a linear-plus-slab benchmark (Shah et al., 2020) with a width-four tanh network, the fixed-scale minimizer was not attained without weight decay (its hidden weights diverged and the slab feature entered gradually); with weight decay it was attained and switched discontinuously toward the slab feature at a single, reproducible output scale. A registered training test of that switch was unresolved on its validity conditions: only 22 of 40 runs crossed (30 were required), and the growth-to-relaxation ratio at crossing (median 7.4) lay far outside that test's registered validity bound of 0.06, so the transfer to this benchmark is not established.

7 Related work

The fixed-scale minimizer is a regularization path. Regularized logistic minimizers converge to maximum-margin separators as regularization vanishes (Rosset et al., 2003). For linear predictors and general decreasing convex losses, gradient descent and the regularization path have the same limiting direction (Ji et al., 2020). Those directional limits leave the present finite-scale nonlinear lag undetermined. The SGD landscape varies with feature-learning strength (Atanasov et al., 2025), while initialization scale relative to training accuracy controls the kernel-to-rich transition in deep linear networks (Moroshko et al., 2020); output-scale analyses also describe lazy training (Chizat et al., 2019), unbalanced feature-learning speed (Kunin et al., 2024) and lazy-to-rich grokking (Kumar et al., 2024; Lyu et al., 2024).

Alternating gradient flows describe when neurons activate and the sizes of resulting loss drops (Kunin et al., 2025). Margin maximization can specify learned features (Morwani et al., 2024), and early simplicity bias can precede margin maximization in two-layer networks (Lyu et al., 2021). Simplicity bias can also select brittle or poorly generalizing features (Shah et al., 2020) and extends beyond linearly separable data (Tsoy & Konstantinov, 2024). Silent alignment separates early kernel alignment from subsequent growth (Atanasov et al., 2022); learning functions of increasing complexity retains an early linear classifier (Kalimeris et al., 2019). Selection by lower-order statistics (Refinetti et al., 2023) is especially close to our small-scale mean preference. Staged dynamics have also been derived for linear networks, matrix sensing and XOR feature learning (Jacot et al., 2021; Pesme & Flammarion, 2023; Jin et al., 2023; Glasgow, 2024).

Slow-manifold tracking and its first-order correction are established concepts (Fenichel, 1979). We add a certified placement change in a non-monotone fixed-scale path, quantitative prospective lag prediction, learning-rate invariance and an empirical validity boundary. Scaling the output weight and bias together scales the network, so the model is quasi-homogeneous in its output layer; margin results for quasi-homogeneous (Kunin et al., 2023) and non-homogeneous (Cai et al., 2025) networks concern the separable limit, whereas the switch here occurs at finite scale, before the relevant branch separates. The lag calculation adds a dynamical test to loss-geometry accounts (Chiang et al., 2023), beyond the existence of low-loss connecting paths (Garipov et al., 2018; Draxler et al., 2018). Our continuous-error check follows the concern that a binary metric can manufacture apparent emergence (Schaeffer et al., 2023); 76% of error improvement occurs while no run solves. Appendix T gives the topological and activation context and the earlier comparators.

8 Methodology and limitations

"Registered" means a prediction and scoring rule committed before data collection, with a repository timestamp rather than an external registry. "Post hoc" means analysis or scoring after data were available; hashing predictions before comparison protects the comparison without making their origin prospective. Failed predictions remain beside the claims they limit. Appendix U groups 33 central failures (31 under registered rules, 2 scored afterwards); the supplementary gives the full census.

Beyond width-one sine, prospective branch selection remains open. The revised Adam measurement rule and the suggested predicted-lag boundary also await prospective support. The loss changes its preferred feature at a computable output scale. Within the measured tracking regime, its local geometry predicts the typical delay before training follows.

REPRODUCIBILITY STATEMENT

The task, loss and precision split are specified in Section 2. Appendices B–I give the training protocols, certificate chain and prospective designs. We distinguish finite-objective certificates from interval-wide classification checks and from statistical uncertainty. Prediction timestamps are repository commits, not an external registry. Frozen predictions precede the scored training runs; the geometry test's excluded timing run is disclosed. Validity-check amendments made before outcomes are recorded separately from verdicts. The registration census preserves failed tests when later analyses succeed. The supplementary material contains the code, the certificates with their SHA-256 manifest (one 205 MB certificate file is supplied as its hash and regeneration command), the independent checker and its tests, and a single command that reruns the ledger and the full test suite.

AI USE STATEMENT

Research ideation and problem formulation. The research question grew out of the cited work on topological expressivity and discussion between the authors. Generative AI tools were used extensively to propose experimental designs, candidate explanations and the predictions tested in this paper, including the conditional-loss analysis, the fixed-scale replays, the scaling reduction and the prospective tests. Each prediction was fixed in a timestamped registration before the corresponding data were collected, and its outcome is reported whether it passed or failed.

Experimental code. Nearly all experimental and analysis code was written by an AI coding agent working in the project repository under the authors' direction. Verification did not rely on the agent's own checks. Six core measurements were reimplemented independently from their mathematical definitions; five agreed (identical classifications, or numerical differences of at most 4.5×10^{-5}) and the sixth differed by one grid step. A ledger verifier recomputes every reported number from committed artifacts and fails if any number lacks a committed producing script. The sweeps were regenerated end to end from a clean checkout, and analytic identities, such as agreement between the margin decomposition and the direct correctness check, were registered as tests. These checks found errors that were corrected before submission, including a hand-transcribed artifact, a random-seed bug that produced an apparent precision effect, and an orientation sign error.

An independent checker re-verifies the certificates and shares no code with the certifying searches. It imports nothing from the project's source tree. It re-implements the profiled loss, class gap and limit problem from their mathematical definitions in python-flint/Arb 80-bit ball arithmetic; the searches used float Lipschitz bounds, numpy outward rounding or mpmath intervals. Every decision is an Arb comparison. For exported certificates it reads only data checked against committed SHA-256 hashes. Its tests include constructed cases it must reject. It verifies the finite-a placement brackets, $\widehat{G}$ enclosures, limit-switch bracket and localization, $K = \sup G_0$ with its domain lemma, first-order Krawczyk boxes, and positive-definite and ring certificates. Outer exclusion and the finite and limiting solve brackets were not independently checked; only the searches support them.

Analysis and interpretation. AI tools were used to analyse results and propose interpretations. Interpretations were tested against registered predictions and direct measurements, and several were withdrawn or narrowed as a result; the text reports these corrections. A co-author reviewed the theoretical claims and identified errors that were corrected. AI assistance was used to check proofs and derivations; the switch locations rest on computer-assisted certificates re-checked by an independent checker.

Writing and literature. The manuscript was drafted and revised with generative AI assistance throughout, under the authors' direction of its structure and claims. Candidate references were proposed with AI assistance, and every included entry was checked against its arXiv, conference or proceedings record. AI was not used to generate datasets, translate text or clean data; the task data are defined analytically.

Verification and responsibility. We have reviewed all AI-assisted work. The authors reviewed the theoretical claims, including the propositions, lemmas and certificate chain, and identified errors that were corrected before submission. The authors verified the experimental results, using the

ledger verifier, which recomputes every reported number from committed artifacts, together with the independent checks described above. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 247–257. PMLR, 2022.

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19540–19569. Curran Associates, Inc., 2023. doi: 10.52202/075280-0858.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Alexander Atanasov, Alexandru Meterez, James Simon, and Cengiz Pehlevan. The optimization landscape of SGD across the feature learning strength. In *International Conference on Learning Representations*, pp. 5740–5779, 2025.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764. Curran Associates, Inc., 2022. doi: 10.52202/068431-1581.

Yongqiang Cai. Achieve the minimum width of neural networks for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6426–6504. PMLR, 2025.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ping-yeh Chiang, Renkun Ni, David Y. Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *International Conference on Learning Representations*, 2023.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318. PMLR, 2018.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940. PMLR, 2016.

Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979. doi: 10.1016/0022-0396(79)90152-9.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *International Conference on Learning Representations*, pp. 52419–52430, 2024.

William H. Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology. arXiv preprint arXiv:1802.04443, 2018.

Boris Hanin and Mark Sellke. Approximating continuous functions by ReLU nets of minimal width. arXiv preprint arXiv:1710.11278, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415, 2016.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. arXiv preprint arXiv:2106.15933, 2021.

Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2109–2136. PMLR, 2020.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15200–15238. PMLR, 2023.

Jesse Johnson. Deep, skinny neural networks are not universal approximators. In *International Conference on Learning Representations (ICLR)*, 2019.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 2020.

Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *International Conference on Learning Representations (ICLR)*, 2024.

Tanishq Kumar, Blake Bordelon, Samuel Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, pp. 23010–23035, 2024.

Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. The asymmetric maximum margin bias of quasi-homogeneous neural networks. In *International Conference on Learning Representations*, 2023.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 81157–81203. Curran Associates, Inc., 2024. doi: 10.52202/079017-2580.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James Simon, Michael Deweese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4377–4424. Curran Associates, Inc., 2025. doi: 10.52202/085713-0156.

Li'Ang Li, Yifei Duan, Guanghua Ji, and Yongqiang Cai. Minimum width of Leaky-ReLU neural networks for uniform universal approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19460–19470. PMLR, 2023.

Ziming Liu, Eric Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12978–12991. Curran Associates, Inc., 2021.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Du, Jason Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, pp. 33897–33936, 2024.

Diganta Misra. Mish: A self regularized non-monotonic activation function. In *31st British Machine Vision Conference (BMVC)*, 2020. doi: 10.5244/C.34.191.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22182–22193. Curran Associates, Inc., 2020.

Depen Morwani, Benjamin Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *International Conference on Learning Representations*, pp. 29077–29114, 2024.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2021.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7475–7505. Curran Associates, Inc., 2023. doi: 10.52202/075280-0329.

Lucas Prieto, Melih Barsbey, Pedro Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *International Conference on Learning Representations*, pp. 81151–81168, 2025.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854. PMLR, 2017.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. arXiv preprint arXiv:1710.05941, 2017.

Maria Refinetti, Alessandro Ingrosso, and Sebastian Goldt. Neural networks trained with SGD learn distributions of increasing complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28843–28863. PMLR, 2023.

Junyu Ren and Lek-Heng Lim. Low-dimensional topology of deep neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, pp. 23881–23904, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581, 2023. doi: 10.52202/075280-2425.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585. Curran Associates, Inc., 2020.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3067–3075. PMLR, 2017.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Matus Telgarsky. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539. PMLR, 2016.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48728–48767. PMLR, 2024.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *International Conference on Learning Representations*, 2023.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1583–1594, 2020.

A Analytic existence of the width-1 switch

The population sample is balanced and contains the window endpoints. Write $u_\theta = \operatorname{sign}(w_2) f_a(w_1 x + b_1)$, $s = |w_2|$, and distinguish the continuous-window gap G from its data-point counterpart G_n. The class-mean identity in the main text follows because the linear and odd sine parts have zero class-mean difference on symmetric data. For continuous uniform windows,

$$d(\alpha) = \frac{\sin(2\alpha) - \sin(1.2\alpha)}{0.8\alpha} - \frac{\sin(0.8\alpha)}{0.8\alpha}.$$

The reference α^* is 1.7913244 on the data in the non-aliased range and 1.7922917 on the continuous windows; the corresponding $|d|$ values are 1.571067 and 1.573242. Identity checks give maximum errors 3.3×10^{-15} and 2.3×10^{-14}, respectively.

Placement bound. Take $w_1 > 0$. In the positive orientation, bound the gap by $f_a(-2w_1 + b_1) - f_a(0.8w_1 + b_1)$; in the negative orientation use $f_a(-0.8w_1 + b_1) - f_a(2w_1 + b_1)$. The sine difference is bounded by $2|\sin(1.4w_1)|$. Reflection handles negative w_1, giving

$$G, G_n \le 2a|\sin(1.4w_1)| - 2.8|w_1| \le 2a - 2.8|w_1|.$$

These points belong to both the windows and the population sample. Thus a placed parameter must have $|w_1| < a/1.4$, whereas this class-mean reference is unplaced for $a < 1.4\alpha^*$. Its gap is -3.66 at $a = 1.30$ and -3.45 at $a = 1.50$.

Theorem 2 (An analytic bracket for the width-1 switch)**.** *Let $1 < a < 1.4\alpha^*$, let θ^* be the mean-gap reference above, and put*

$$d_P(a) = \max_{|\alpha| \le a/1.4} |d(\alpha)|, \quad s_0 = \frac{2(\Delta\mu(\theta^*) - a d_P)}{\mathrm{Var}(u_{\theta^*})}, \quad s_1 = \frac{2\log(n/\log 2)}{\widehat{G}(a) - \eta}.$$

For $s < s_0$, every placed parameter has higher profiled loss than the unplaced θ^*. When $\widehat{G} > \eta$, for $s > s_1$, every unplaced parameter has higher loss than the placed witness. Here $\eta = 0$ for data placement; for window placement use the grid correction with $\widehat{G} > \eta$.

Proof of Theorem 2. For labels $y \in \{0, 1\}$, convexity gives $\ell(z, y) \ge \log 2 + (1/2 - y)z$. Balance cancels the bias, so $\mathcal{L}^*(\theta; s) \ge \log 2 - s\Delta\mu(\theta)/4$. Every placed parameter has $\Delta\mu \le a d_P$. At θ^*, the bound $\ell'' \le 1/4$ and bias $b = -s\mathbb{E}u_{\theta^*}$ give

$$\mathcal{L}^*(\theta^*; s) \le \log 2 - \frac{s}{4}\Delta\mu(\theta^*) + \frac{s^2}{8}\mathrm{Var}(u_{\theta^*}).$$

Comparison proves the small-scale statement for $s < s_0$.

For $G_n(\theta) \le \eta$, the extreme data points give $\mathcal{L}^*(\theta; s) \ge (\log 2/n)e^{-s\eta/2}$. The witness has $G_n \ge G \ge \widehat{G}$ and loss less than $e^{-s\widehat{G}/2}$. Comparison proves the large-scale statement for $s > s_1$.

For continuous-window placement, put $\eta = (a/1.4)(1 + a)(h_I + h_O)$, where h_I, h_O are grid half-spacings. If $G \le 0$ and $|w_1| < a/1.4$, Lipschitz continuity gives $G_n \le G + |w_1|(1+a)(h_I + h_O) \le \eta$. If $|w_1| \ge a/1.4$, the endpoint bound already gives $G_n \le 0$. This proves the window version when $\widehat{G} > \eta$. The proof compares all candidates with explicit witnesses and requires neither attainment nor compactness. □

The constants use $d(\alpha^*)$, the restricted maximum d_P (with a Lipschitz grid allowance), the variance and the rigorous geometry witness. The window correction is admissible at every tabulated activation value except $a = 1.02$. The bracket is deliberately loose: it proves existence, while the certificates locate the switch.

Table 3: Analytic data-placement bounds and certified threshold intervals on the population objective. These are deterministic bounds, not statistical intervals.

a	s_0	s_1	Certified $\lvert w_2\rvert_{\mathrm{glob}}$
1.30	0.342	163.8	$[4.95, 4.963]$
1.35	0.327	131.9	$[4.025, 4.037]$
1.40	0.310	109.5	$[3.375, 3.388]$
1.45	0.293	93.1	$[2.888, 2.9]$
1.50	0.274	80.5	$[2.525, 2.538]$
1.60	0.236	62.9	$[2, 2.013]$

A.1 Output-scale law and the bound on Adam's step

Equation (5) follows from $A = s\varepsilon^{3/2}$ and the first-order calculation in Appendix F.2. The independently checked bracket is $A^* \in [0.68125, 0.6875]$; the Krawczyk value is 0.6854452. The relative first-order correction has enclosure $[0.6621547, 0.6621550]$. Its finite-a errors range from $+0.9\%$ at $a = 1.30$ to $+2.7\%$ at $a = 1.60$; without it the limiting approximation is 16–27% low.

For Adam with learning rate $\eta_{\text{opt}} = 0.01$, $(\beta_1, \beta_2) = (0.9, 0.999)$ and denominator constant 10^{-8}, Cauchy–Schwarz gives, for any gradient sequence,

$$|\Delta\theta_i| \le \eta_{\text{opt}} B_t, \qquad B_t = \frac{1-\beta_1}{1-\beta_1^t}\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}\sqrt{\frac{1-\zeta^t}{1-\zeta}}, \quad \zeta = \frac{\beta_1^2}{\beta_2}.$$

Apply the inequality to $\sum_k \beta_1^{t-k} g_k$ using weights β_2^{t-k}. The resulting bound satisfies $B_t \le B_\infty = 7.2703$. A geometric gradient sequence $g_k \propto (\beta_1/\beta_2)^{t-k}$ attains the Cauchy–Schwarz ratio; the positive denominator constant can only reduce the actual step. Summing bias-corrected bounds gives $|w_2| \le 106.93$ after 2,000 steps from $|w_2(0)| \le 1$.

At $a = 1.02$, the scale law gives 245.6, with 240.9 even at the certified lower end of A^*. The threshold-crossing picture therefore requires at least 3,966 steps under the worst-case move bound. The observed terminal scale at 2,000 steps has median 1.85 and maximum 3.92, with zero of 200 runs solving. This is consistent with, and forced by, the threshold-crossing picture; it is not independent evidence for that picture, nor an impossibility theorem for all correct classifiers below the conditional threshold.

The bound explains divergence of the required scale as $\varepsilon^{-3/2}$ and of the time needed to reach it under bounded steps. It does not derive Adam's growth. At 2,000 steps it permits $a \ge 1.035$, while the observed onset is 1.60; at 128,000 it permits $a \ge 1.0018$, while the observed onset is 1.03. The budget law combines the derived geometric exponent with measured growth exponent $\alpha = 1.1172$. Its derivation failed under SGD. Neither this bound nor that effective exponent explains the lag; Appendix V gives the tracking calculation and its subsequent prospective test.

A.2 Hypothesis audit and the finite-grid alias

For the sine family, positive data gap implies $|w_1| < a/1.4$. Periodicity in b_1 and continuity therefore give attainment of $\Gamma_n = \sup G_n \ge \widehat{G} > 0$. The large-scale comparison in Theorem 2 forces conditional minimizers into this compact set, modulo symmetries. At small scale, $W(s)$ is finite for each $s > 0$ but diverges as $s \to 0$; a common compact set is not proved. At $a = 1.30$, $sW(s) \to 4.51$.

The uniform remainder is proved by Lemma 4; continuity and boundedness on the common compact set make its bound uniform. Quadratic growth is no longer required for Theorem 1. The earlier numerical checks are retained as a record. Numerical checks at sampled parameters and scales give maximum $|\text{remainder}|/(s^4 m_4) = 0.0052084$; this does not certify that constant globally. Float64 Lipschitz-grid checks give $|D''| \ge 2.454$ on the data and 2.455 on continuous windows near α^*, yielding $|D^*| - |D(\alpha)| \ge 1.227(\alpha - \alpha^*)^2$. These are numerical checks, not interval certificates. Quadratic growth near the aliased maximizers below was not examined.

On the continuous windows the checked global maximizer is $\alpha_c^* = 1.7922917$. On the 800-point quadrature, every point is an integer multiple of $q = 0.4/79401$, with even inner integers and odd outer integers. Thus $|D(\pi/q)| = 2 > 1.571067$: the data mean-gap supremum is attained at an alias, not at $\alpha^* = 1.7913244$. The latter is maximal only on $(0, 622047)$. The alias beats the reference only below output scales 2.0×10^{-12} to 2.5×10^{-12} over $a = 1.30$–1.60; localization excludes $|w_1| \ge \pi/q$ for $s > 7.24 \times 10^{-6}$ at $a = 1.30$. Both candidates are unplaced, leaving the width-one conclusion unchanged. The explicit small-scale bracket only needs the reference to beat all placed candidates, not to maximize the mean gap globally. The width-two asymptotic selection statement is restricted to continuous windows; its reported finite-scale checks still use continuous-window extrema.

B Training protocols and correctness

Every experiment that replays several arms from a saved optimizer state was checked arm by arm against its checkpoint, bit for bit. All were clean apart from the two lag-intervention arms detected by the audit; their corrected reruns are disclosed in Appendix H.

The base width-1 protocol uses 200 uniform points per class, full-batch updates, initialization $U(-1,1)^4$ and mean logistic loss

$$\mathcal{L} = \frac{1}{n}\sum_i [\log(1+e^{N(x_i)}) - y_i N(x_i)],$$

with $y_i = 0$ on I and $y_i = 1$ on O. Adam uses learning rate 10^{-2} and a default budget of 2,000 steps. AdamW uses the same rate and weight decay 0.01; SGD uses rate 0.3. The new own-sample prospective design instead uses 200 total training points per run. Sample-size and budget tests vary their named quantities.

Headline sweeps use float32; conditional analyses, barrier checks and Hessian calculations use float64. Shared initializations give no solve-status flips in 180 paired precision checks. Drawing separate random initializations in each dtype is not such a paired check. Sweep evaluation uses 4,001 inner and $2 \times 2{,}000$ outer grid points. This differs from the registered fresh-point evaluation and is recorded as a protocol deviation. Region-wide certification is separate: all 870 solved runs in the pooled float32 and float64 population are certified, whereas Phase 1's 430 solved runs belong to float32 only.

In the earlier trajectory sweep, placement-crossing steps range from 1,860 to 6,839 at $a = 1.30$ and budget 8,000 only. New threshold figures use the 32,000-step cohort. The earlier growth-rate measurements show no acceleration at placement crossing, but the new phase-specific interventions are needed to interpret the lag.

C Margin identity, bound and verification

Theorem 3 (Margin bound)**.** *A classifier with logit margin $m > 0$ has $G > 0$ and $s \geq 2m/G \geq 2m/G^\star(a)$.*

For $s = w_2 > 0$, the worst logits in the two classes are $sV + b_2$ and $sU + b_2$. Therefore

$$\gamma = \min(sV + b_2, -sU - b_2) = \frac{s(V-U)}{2} - \left| b_2 + \frac{s(U+V)}{2} \right|.$$

The conditions for margin m are $sV + b_2 \geq m$ and $sU + b_2 \leq -m$. Subtraction gives $sG \geq 2m$, proving Theorem 3. For any placement with $G > 0$, setting $b_2 = -s(U+V)/2$ constructs a correct classifier with margin $sG/2$. Taking a supremum over placements gives the supremum of achievable margins $sG^\star/2$, even if it is not attained. The odd-activation reparameterization in Section 2 proves the negative-orientation case.

The per-placement inequality was checked on 66 solvers recorded before its derivation, including 22 with negative output weight. There were no violations; median slack was 1.2488 and minimum slack 1.0070. This checks the bound at each solver's actual placement. It does not require identifying the global gap supremum. Uniform numerical verification must use the certified upper endpoint G_+ in $2m/G_+$; enlarging a numerical gap estimate weakens that lower bound. Refining a grid alone cannot certify the optimum.

At larger width, one must analyze the sum of hidden-unit outputs. In particular, a sum of per-unit worst-case gaps is generally a lower bound on the gap of the sum, so the width-1 proof does not imply a necessary inequality obtained by summing those individual gaps. No width-3 or width-5 claim here rests on such an extension.

D Exact extrema and certified geometry

The dip depth is

$$D(a) = 2\big(\sqrt{a^2-1} - \arccos(1/a)\big) \sim (4\sqrt{2}/3)\varepsilon^{3/2}. \tag{8}$$

For $w_1 \neq 0$, an extremum of $f_a(w_1 x + b_1)$ on a class interval occurs at an endpoint or at

$$w_1 x + b_1 = \pm \arccos(-1/a) + 2\pi k, \qquad k \in \mathbb{Z},$$

when the corresponding x lies in that interval. Evaluating this finite set gives U and V. When $w_1 = 0$, the hidden output is constant and the gap is zero. The procedure checks whole intervals rather than sample points. Region-wide Lipschitz certification covers all 870 solved runs pooled across the float32 and float64 populations: 866 are certified at grid spacing 4×10^{-4} and the remaining four at 4×10^{-5}. This certification population differs from the float32 Phase 1 population in Section 2.

The placement search can be reduced to a compact domain analytically. Write $i = 0.8$ and $o = 2.0$. If $w_1 > 0$, the pair $x = -o \in O$, $x = i \in I$ gives

$$G(w_1, b_1) \le f_a(b_1 - ow_1) - f_a(b_1 + iw_1) \le -(o + i)w_1 + 2a.$$

For $w_1 < 0$, use $x = o$ and $x = -i$ to obtain the same bound in $|w_1|$. Thus any positive gap has $|w_1| < 2a/(o + i)$. Also $f_a(t + 2\pi) = f_a(t) + 2\pi$, so shifting b_1 by a period leaves G unchanged. The compact rectangle $|w_1| \le 2a/(o + i)$, $b_1 \in [0, 2\pi]$ therefore contains a maximizer whenever a positive gap exists. This establishes compactness independently of observations on a boundary shell.

The gap also satisfies $G \le D(a)$. In the pair just used, the outer preactivation precedes the inner one. The largest downward excursion of f_a is its local maximum-to-minimum drop $D(a)$: successive local maxima and successive local minima each increase by 2π, so extending across further periods cannot increase the drop. Hence $G^\star/D \le 1$.

Exact class extrema remove the inner sampling error. For certification we maximize both output orientations, using $\mathcal{G}(u) = \max\{\min_O u - \max_I u, \min_I u - \max_O u\}$. The opposite orientation obeys the same compact-domain bound by comparing the other pair of window endpoints. Symmetry reduces the search to $w_1 \in [0, a/1.4]$, $b_1 \in [0, 2\pi]$, with zero gap at $w_1 = 0$.

The rebuilt certificate uses branch and bound on the larger rectangle $[0, a] \times [0, 2\pi]$. For a cell with half-widths (h_w, h_b), its upper bound is

$$\mathcal{G}(\text{center}) + (1 + a)(2.8h_w + 2h_b).$$

The coefficient 2.8 adds the inner and outer bounds on $|x|$; both extrema contribute to the bias term. Cells start with half-widths 0.05, halved each round. A cell is discarded when its upper bound is no larger than the best attained gap. The search stops when the largest surviving upper bound is within 10^{-3} relative to that gap. All 13 activation values stop within 10–19 rounds.

The independent checker replaces the original float endpoints with rigorous enclosures. We retain the same witness for $\widehat{G}$ in R, using the checker's downward-rounded lower bound there. The best certified lower bound G_- can come from a different witness. The largest relative difference $G_-/\widehat{G} - 1$ is 8.36×10^{-5} at $a = 1.10$, or 1.48×10^{-6} in absolute gap; the largest absolute difference is 7.09×10^{-6} at $a = 1.60$. The corresponding enclosure for $G^\star/D$ is $[G_-/D, G_+/D]$.

The rigorous replacement changes $\widehat{G}$ by at most 8.4×10^{-14} relative. All 392 printed-number checks remain stable, including stored R values computed before the replacement. The original float endpoints can miss the rigorous enclosure by 1.1×10^{-15}; they are not the verified endpoints.

The geometric constant is task-dependent. Rescaling every window edge is absorbed by the input weight, whereas changing window ratios changes the optimization over placements. This is why a separate limiting K is computed for each geometry.

E Conditional certificates and search audit

The conditional objective is mean logistic loss on 800 deterministic quadrature points, with 400 on the inner interval and 200 on each outer interval. Its profiled bias is the unique solution of

$$\frac{1}{n}\sum_i \text{sigmoid}\big(sf_a(w_1x_i + b_1) + b_2\big) = \frac{1}{n}\sum_i y_i.$$

The left side is strictly increasing in b_2. Bias profiling is therefore an exact reduction of the optimization problem; interval enclosures bound its numerical evaluation. Period shifts of b_1 are absorbed in b_2. The finite-a certificate restricts $|w_1| \le W(s, a)$ using an analytic bound, then certifies the profiled problem by branch and bound. The loss certificate and the maximal-gap certificate concern different objectives.

Table 4: Rigorous geometry values from the independent checker. R uses $\widehat{G}$, the lower bound at its retained witness; $[G_-, G_+]$ separately encloses $G^\star$ using the best certified lower bound. These are deterministic enclosures, not confidence intervals.

a	$\widehat{G}$ used in R	G_-	G_+
1.02	0.001626744762	0.001626745002	0.001627653245
1.05	0.006360085994	0.006360085994	0.006363833420
1.10	0.017671950853	0.017673427712	0.017688791467
1.15	0.031918407628	0.031920435133	0.031951889476
1.25	0.066505622211	0.066507277337	0.066573122000
1.30	0.086101793878	0.086103036064	0.086170343941
1.35	0.106913550661	0.106914064287	0.106982835379
1.40	0.128771895556	0.128772506040	0.128842740347
1.45	0.151545318737	0.151545622037	0.151689017080
1.50	0.175126200489	0.175126200489	0.175272043337
1.60	0.224360202126	0.224367291384	0.224519465715
2.00	0.441033924063	0.441033924063	0.441384755802
3.00	1.052297757851	1.052297757851	1.053233148730

Lemma 1 (Finite-a localization)**.** *On a balanced sample from the base windows, orient $s = |w_2| > 0$ and write $z(x) = s f_a(w_1 x + b_1) + b_2$. For a cut $c \in [0, 0.8)$ define*

$$\pi_+(c) = \frac{\min\{\#(O \cap (-\infty, -1.2]),\ \#(I \cap [c, \infty))\}}{n}, \quad \delta(\pi) = 2\log(2^{1/\pi} - 1).$$

Let $\mathcal{C} = \{0.799k/79 : k = 0, \ldots, 79\}$ and

$$W_+(s, a) = \min_{c \in \mathcal{C}} \frac{2a + \delta(\pi_+(c))/s}{1.2 + c}.$$

The reflected counts define W_-; cuts with zero count provide no finite bound. If $w_1 > W_+$, the profiled loss exceeds $\log 2$ for every b_1. The reflected statement holds for $w_1 < -W_-$. Thus every global conditional minimizer has $|w_1| \le W(s, a) = \max(W_+, W_-)$, and b_1 can be restricted to $[0, 2\pi)$.

Proof. Since $t - a \le f_a(t) \le t + a$, for $w_1 > 0$ every left-outer logit is at most $A = s(-1.2w_1 + b_1 + a) + b_2$, and every inner logit with $x \ge c$ is at least $A + \Delta$, where $\Delta = s((1.2 + c)w_1 - 2a)$. If $A + \Delta/2 \ge 0$, those inner class-0 points incur loss at least $\mathrm{softplus}(\Delta/2)$. Otherwise those outer class-1 points incur loss greater than that value. Other losses are nonnegative, so profiling the bias preserves

$$\mathcal{L}^*(w_1, b_1; s) \ge \pi_+(c)\,\mathrm{softplus}(\Delta/2).$$

This exceeds $\log 2$ when $\Delta > \delta(\pi_+(c))$, giving the stated bound for every cut and hence their minimum. Reflection gives W_-. The constant predictor attains $\log 2$, so no global minimizer lies beyond either bound. Finally, a period shift of b_1 adds $2\pi s$ to all logits, which b_2 absorbs. □

The quadrature uses 400 equally spaced inner points and 200 equally spaced points on each outer interval. At $a = 1.30$, $s = 4.95$, the grid-optimal cut is $0.799 \cdot 22/79 = 0.22251$ at the reported precision, with 145 inner points, $\pi = 0.18125$ and $\delta = 7.6045$. Both orientations give $W = 2.908$. The implementation agrees with the formula to 10^{-12} at both ends of every reported threshold bracket. Table 5 gives the lower endpoints. The sampled loss checks support the implementation; the lemma supplies the global exclusion.

The retained candidate's gradient norm is at most 2.3×10^{-6}. Replaying the frozen search at all 358 scales reproduces 358 of 358 results. Among 16,468 candidates, none discarded by a screening rule has lower loss, including the exact constant predictor. A stricter optimization with ten times as many steps and a 10^{-8} gradient stopping tolerance agrees on the 12 threshold losses to 2×10^{-16}. A full scan at $a = 1.30$ covers output weights from 1.5 to 11 in steps of 0.5; the minimizing branch is continuous and its gap changes sign once between 4.5 and 5.0. These audit checks met their registered decision rule.

Table 5: Finite-a localization at lower bracket endpoints. The last column is the minimum evaluated profiled loss at $|w_1| = W, 1.5W, 3W$, on 720 biases and both orientations; it is a sampled check, not a certificate.

a	Bracket	s	W	Cut	Inner count	π	Sampled loss
1.30	glob	4.9500	2.908	0.2225	145	0.18125	6.652
1.30	solve	7.1000	2.562	0.3135	122	0.15250	7.040
1.35	glob	4.0250	3.226	0.2225	145	0.18125	6.397
1.35	solve	5.7125	2.828	0.2731	132	0.16500	7.170
1.40	glob	3.3750	3.548	0.1821	155	0.19375	5.490
1.40	solve	4.7375	3.097	0.2225	145	0.18125	7.256
1.45	glob	2.8875	3.876	0.1618	160	0.20000	4.885
1.45	solve	4.0125	3.371	0.2225	145	0.18125	6.416
1.50	glob	2.5250	4.199	0.1416	165	0.20625	4.543
1.50	solve	3.4750	3.646	0.2023	150	0.18750	5.667
1.60	glob	2.0000	4.861	0.1214	170	0.21250	4.337
1.60	solve	2.7000	4.218	0.1618	160	0.20000	4.822

Table 6: Certified finite-a conditional thresholds on the quadrature objective. The $\widehat{G}$ column gives the rigorous lower bound at the retained witness used in R. Every threshold is an enclosure, not a frozen-grid point estimate.

a	$\widehat{G}$	R_{glob}	R_{solve}
1.30	0.086101793878	$(0.21310, 0.21364]$	$(0.30566, 0.30620]$
1.35	0.106913550661	$(0.21516, 0.21583]$	$(0.30537, 0.30604]$
1.40	0.128771895556	$(0.21730, 0.21811]$	$(0.30503, 0.30583]$
1.45	0.151545318737	$(0.21879, 0.21974]$	$(0.30404, 0.30498]$
1.50	0.175126200489	$(0.22110, 0.22219]$	$(0.30428, 0.30538]$
1.60	0.224360202126	$(0.22436, 0.22576]$	$(0.30289, 0.30429]$

The output-weight brackets have width 0.0125. At each solve bracket's lower and upper endpoints, interval evaluation over the global-minimizer enclosure certifies negative and positive worst-case margin, respectively, with the bias validated. At the placement brackets, competing sign-region loss intervals are disjoint. The rounding audit re-evaluates attained losses in interval arithmetic; the reported discrepancy is at most 1.1×10^{-16}. Certificate separation is checked against a rounding allowance, not inferred from agreement between optimization restarts.

Population and sample objectives. Fifty own-sample certificates at each of $a = 1.30, 1.45, 1.60$, with 400 training points each, have median threshold 1.06 times the quadrature threshold, interquartile range 0.99–1.12 and range 0.87–1.22. The quadrature value lies at the 30th–32nd percentile. Evaluations unresolved at tolerance 10^{-9} are treated as unplaced. This affects 20, 18 and 6 of the 50 seeds at the respective activation values, biasing their thresholds upward by at most one bisection step: 0.08%, 0.14% and 0.19%. Every sample has one switch.

E.1 Independent certificate verification

Certificates are exported as data and verified by an independent checker, written separately from the searches, using python-flint/Arb 80-bit ball arithmetic. It imports no search code and makes each verification decision by a ball-arithmetic comparison.

The completed independent checks are:

- The finite-a placement-threshold brackets for $a = 1.30, 1.35, 1.40, 1.45, 1.50, 1.60$ are independently verified in ball arithmetic. All 12 endpoint certificates pass every leaf claim and exact tiling check; the checker recomputes W and matches the hashes.
- The rigorous geometry enclosures for $a = 1.02$, 1.05, 1.10, 1.15, 1.25, 1.30, 1.35, 1.40, 1.45, 1.50, 1.60, 2.00, 3.00 are independently verified. All 13 certificates pass the hash, ex-

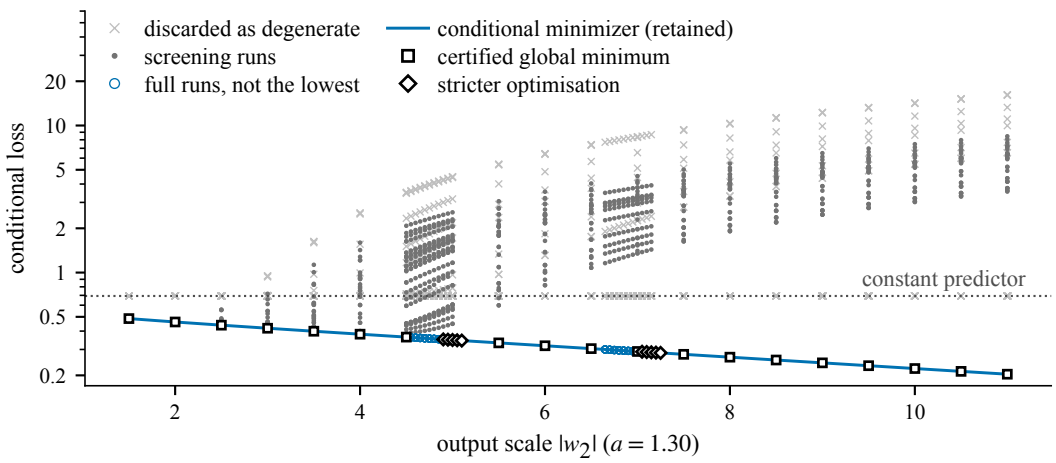


Figure 6: No evaluated search candidate improves on the retained minimizer. Unlike the gap curve in Figure 7, this plot shows candidate losses at $a = 1.30$: 2,701 candidates at 59 evaluations, including 1,194 discarded as degenerate, 976 screening runs and 413 full runs that were not lowest and lie on the retained minimizer. The frozen result reproduces at 59 of 59 evaluations. The plot includes 20 certified global minima and 10 stricter-optimization points. No statistical uncertainty is shown.

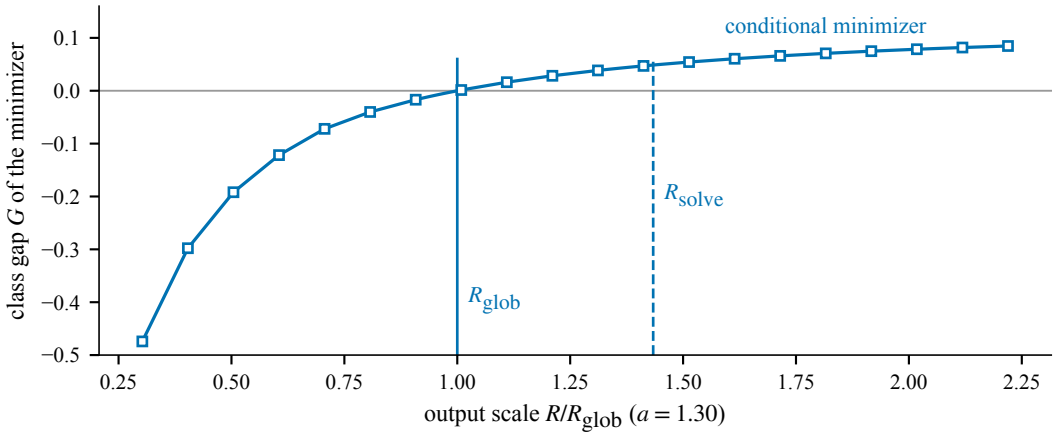


Figure 7: The conditional minimizer's class gap changes sign once, at R_{glob}. Unlike the threshold comparison in Figure 3, this plot resolves the gap along the minimizing branch. At $a = 1.30$, the 800-point population objective is minimized at 39 output scales (line), with certified global minimizers at 20 scales (squares). No uncertainty is drawn; certified losses are accurate to at most 10^{-9}. The horizontal coordinate is $R/R_{\text{glob}} = |w_2|/4.95625$. Placement and solve brackets are $|w_2| \in (4.9500, 4.9625]$ and $(7.1000, 7.1125]$, giving $R_{\text{solve}}/R_{\text{glob}} = 1.434$. Minimizers on opposite sides of the switch are at most 2.7×10^{-4} apart; other basins lie at least 0.0031 higher in loss outside radius 0.1, both certified.

act domain-tiling and analytic domain-reduction checks. The rigorous endpoints in Table 4 replace the original float endpoints.

- The limit-switch bracket $A^* \in [0.68125, 0.6875]$ is independently verified at both ends. These checks cover every losing-region leaf, exact tiling and data symmetry. The localization bounds $B(24)$ and B_{full} were recomputed exactly and verified within these limit-switch checks.

The additional independent checks are:

- $K = \sup G_0$ over $[0, 8] \times [-12, 12]$: fresh Arb branch and bound verifies the upper bounds 0.5794951 and 0.5794559217, the witness lower end, and the domain lemma in exact rational arithmetic.
- The first-order Krawczyk boxes: the switch has a unique zero in its 10^{-9} box, $A^* \in [0.6854452375756532, 0.6854452375756537]$ and $A'(0)/A^* \in [0.6621547824344876, 0.6621549498240902]$, with unique active set. The cubic-gap vertex check also passes.
- The positive-definite boxes and no-critical-point rings pass for both placement and solve chains. The Hessian margins are 0.0473 and 0.0209, respectively; the checker re-brackets the profiled bias.

An independent checker in ball arithmetic (Arb) re-verifies the finite-a placement brackets, the $\widehat{G}(a)$ enclosures, the limit switch bracket and its localization, $K = \sup G_0$ with its domain lemma, the Krawczyk boxes of the first-order calculation, and the positive-definite and ring certificates of the uniqueness chain. The outer-exclusion certificates and the solve brackets are verified by the original searches only.

The independently checked A^* and K enclosures give $R^{\infty}_{\mathrm{glob}} \in [0.19738, 0.19920]$. The sharper interval $[0.1985926, 0.1985927]$ needs the Krawczyk zero to be the global switch, which uses outer exclusion. It is therefore certified but not yet fully independently verified. Independent outer exclusion and finite and limiting solve checks were not run because of their projected cost; no verification is claimed for them.

All ten added first-order endpoint certificates pass independent ball-arithmetic checks. The rescaled geometry enclosures remain search-certified only (Appendix F.2).

Timing the checker exposed three bugs in it, including one where a ball containing zero could produce NaN; all were fixed and tested before the reported checks, and an earlier check affected by them was voided and rerun.

The supplementary contains the independent checker and its tests, a SHA-256 manifest with commands regenerating every certificate bit-identically, the exporting searches, and a single command that reruns every reported check with its expected output. Certificate data are regenerated and checked against their hashes rather than included directly.

F SCALING PROOF AND CERTIFICATE CHAIN

The objects and hypotheses. Let $G_0(p, q) = \min_O h(px + q) - \max_I h(px + q)$ in the positive output orientation. The gap-maximizing placement defines $K = \sup G_0$. It need not minimize logistic loss. The latter has a local minimizing branch and a globally preferred solution; identifying them requires competitor exclusion.

H1 localizes the gap maximizer: for sufficiently small positive ε, it is attained in

$$\mathcal{C} = \{(p, q) : 1 \le |p| \le 2.2,\ |q| \le 2\},$$

up to period shifts and the symmetries $\sigma \mapsto -\sigma$ and $x \mapsto -x$. Lemma 3 proves H1 for $0 < \varepsilon \le 0.029$, supporting the scaling of $G^\star$. It is distinct from H1′, localization of the limiting loss minimizer, discharged by the certificate below. H2′ is a nonzero derivative of the branch's gap at its zero, together with an invertible Hessian and a locally fixed active extremal pair. The certificate supplies positive definiteness and transversality for the base quadrature objective. The finite-a local corollary does not assume global optimality of its continuation.

Lemma 2 (Domain of the cubic gap)**.** *Let $h(t) = -t + t^3/6$ and $G_0(u, v) = \min_O h(ux + v) - \max_I h(ux + v)$. For $u \ge 0$, both this gap and the reverse-orientation gap are nonpositive if*

$$u \ge \sqrt{50/3} \qquad \text{or} \qquad |v| \ge \sqrt{2} + 0.6u.$$

Since $G_0(-u, v) = G_0(u, v)$ and $K > 0$, the unrestricted supremum satisfies

$$K = \sup_{(u,v)\in\mathbb{R}^2} G_0(u, v) = \max_{(u,v)\in[0,8]\times[-12,12]} G_0(u, v).$$

Proof. For $d \geq 0$,

$$h(t) - h(t+d) = d\left[1 - \frac{(t+d/2)^2 + d^2/12}{2}\right].$$

Write $\sigma(x) = ux + v$. For the positive output orientation, the gap is at most either $h(\sigma(-2)) - h(\sigma(-0.8))$ or $h(\sigma(-2)) - h(\sigma(0.8))$. The respective differences have $d = 1.2u$ and $d = 2.8u$; the second has midpoint $v - 0.6u$. In the reverse orientation, use $h(\sigma(0.8)) - h(\sigma(2))$ and $h(\sigma(-0.8)) - h(\sigma(2))$, whose midpoint in the second difference is $v + 0.6u$.

If $u \geq \sqrt{50/3}$, the first difference in either orientation is nonpositive because $d^2/12 \geq 2$. If $|v| \geq \sqrt{2} + 0.6u$, the second difference in either orientation is nonpositive because its squared midpoint is at least 2. Symmetry of both windows under $x \mapsto -x$ gives $G_0(-u, v) = G_0(u, v)$. Every positive-gap placement therefore lies, after this reflection, inside $0 \leq u < \sqrt{50/3}$ and $|v| < \sqrt{2} + 0.6\sqrt{50/3}$, a subset of the searched box. The positive witness and continuity give the claimed equality and attainment. □

Expansion and loss reduction. Taylor expansion on a bounded rescaled domain gives

$$\phi_\varepsilon(\sigma) = h(\sigma) + \varepsilon r(\sigma) + O(\varepsilon^2), \qquad h = -\sigma + \sigma^3/6, \qquad r = \sigma^3/6 - \sigma^5/120.$$

The $(1 + \varepsilon)$ factor in the sine activation contributes the cubic term in r. Each extremum changes by at most the uniform function error, so the gap functional is 2-Lipschitz. The proved H1 then gives $G^\star = K\varepsilon^{3/2}(1 + O(\varepsilon))$. With shifted bias $b = b_2 + s\pi$, Here $\widetilde{R} = sG^\star/2$ is the exact-gap normalization, distinct from the conditional threshold R^*.

$$\mathcal{L}_\varepsilon(p, q, b; \widetilde{R}) = \frac{1}{n}\sum_i \ell\left(\frac{2\widetilde{R}}{K} h(px_i + q) + b, y_i\right) + O(\widetilde{R}\varepsilon).$$

Logistic loss is 1-Lipschitz in the logit. Profiling b preserves the loss-value bound. On the neighborhood used for the implicit-function argument, the expansion also holds in C^2.

Localization of the limit loss. For $h(\sigma) = -\sigma + \sigma^3/6$, removing the points with $|px + q| < 2\sqrt{2}$ leaves a monotone logit. If $|p| \geq P$, removed points lie in an input interval of length at most $4\sqrt{2}/P$. The minimum isotonic logistic loss over the remaining points is a lower bound for the original loss. Minimizing this bound over interval positions gives $B(P)$. If $|q| > 2\sqrt{2} + 2|p|$, the full sample has a monotone logit.

For the fixed quadrature objective, exact isotonic block sums give

$$B(24) = 0.38797358, \qquad B_{\text{full}} = \log(3)/4 + \log(3/2)/2 = 0.47738563.$$

A candidate branch has loss at most 0.3768942 throughout $A \in [0.66, 0.72]$. It beats both bounds. Hence every global minimizer lies in $\mathcal{K}(24)$. Uniformity is established by convexity of the profiled loss in A at fixed (p, q), not by sampling A alone. The solve bracket has candidate upper bound 0.28117, also below $B(24)$.

Global uniqueness through the switch. The certificate covers every $A \in [0.66, 0.71]$, carrying A as an interval through 40 sub-intervals. It has four links:

1. Localization puts all global minimizers in $\mathcal{K}(24)$.
2. Outside a radius-0.15 box around $(p_0, q_0) = (1.66858, 1.36972)$, the certified loss exceeds the candidate upper bound by at least 1.00×10^{-4}.
3. On the ring between radii 0.05 and 0.15, interval gradients exclude a critical point. The smallest certified nonzero gradient component is 3.28×10^{-3}.
4. On the inner box, interval bias enclosures and the full (p, q, b) Hessian give $\lambda_{\min} \geq 0.0473$ throughout the A range.

A global minimizer exists, lies inside the outer box and must be a critical point. The ring excludes it, and positive definiteness makes the inner critical point unique. Reflection supplies the equivalent

branch at negative p. This identifies the global minimizer with the branch continued by the implicit-function theorem.

The outer bound includes a 10^{-9} rounding allowance. The ring uses outward rounding after every arithmetic operation, with widened transcendental evaluations and a summation error bound. The bias root is enclosed using certified signs of monotone sigmoid averages. Positive-definiteness checks enclose the Hessian on 500 sub-boxes. Constructed invalid cases, including a displaced center and an insufficient domain, fail certification. These checks test the certificate implementation; they are not experiments supporting the placement hypothesis.

Switch and transversality. The minimizing gap is certified negative at $A = 0.68125$ and positive at $A = 0.6875$. On its smooth active pair, outer $x = -1.2$ and inner $x = 0.8$,

$$\frac{dG_0}{dA} = \nabla G_0 \cdot \frac{d\theta}{dA}, \qquad \frac{d\theta}{dA} = -H^{-1}\partial_A \nabla \mathcal{L}_0.$$

Interval evaluation gives $dG_0/dA \in [1.666, 1.705]$ over the switch enclosure. A Krawczyk certificate for $(\nabla_\theta \mathcal{L}_0, G_0) = 0$ sharpens the root to $A^* = 0.68544523757565$ at the reported precision. Combining its enclosure with the global gap certificate and Lemma 2, using the outer-exclusion link that has not been independently checked, gives $R^\infty_{\mathrm{glob}} \in [0.1985926, 0.1985927]$.

The same search-certified four-link argument, whose outer exclusion and solve bracket have not been independently checked, establishes global uniqueness on the solve bracket $A \in [1.05875, 1.06]$. There the outer exclusion margin is at least 1.48×10^{-4}, the ring gradient bound is 8.47×10^{-3} and the neighborhood Hessian bound is 0.0209. The margin crosses transversally, with derivative in $[0.532, 0.536]$, giving $R^\infty_{\mathrm{solve}} \in [0.30675, 0.30712]$ under the global cubic-gap normalization.

Finite-a continuation. On the inner box, $|\sigma| \le 4.857$. For $\varepsilon \le 0.05$, the derivative bounds

$$|\phi^{(k)}_\varepsilon - h^{(k)}| \le \varepsilon M_k, \qquad (M_0, M_1, M_2, M_3) = (5.217, 13.507, 16.375, 13.384)$$

give uniform C^2 convergence of the loss. For sufficiently small ε, the Hessian remains positive definite and the local minimizer persists. The gap derivative converges under the fixed active pair, so its transversal zero moves by $O(\varepsilon)$. No explicit ε_0 is claimed. The localization proof uses the cubic's monotonicity outside its central dip and does not transfer to a sine activation with repeated dips. Finite-a global preference therefore rests on the separate certificates in Table 6.

F.1 Proof of gap localization for small positive epsilon

Here $K(\varepsilon) = G^\star(1+\varepsilon)/\varepsilon^{3/2}$ uses the true supremum. We allow both output orientations; their supremum agrees with the positive-orientation one by oddness. No numerical certificate is used in this localization proof.

Lemma 3 (H1 for $0 < \varepsilon \le 0.029$)**.** *The positive maximal gap is attained. Every maximizer has, up to period shifts and reflections, rescaled coordinates satisfying*

$$1.0155 \le K(\varepsilon)/0.4 \le p < \sqrt{6}/1.4, \qquad |q| < 1.7956.$$

It therefore lies in $\mathcal{C} = \{1 \le |p| \le 2.2,\ |q| \le 2\}$, and $|K(\varepsilon) - K| \le 23.3\varepsilon$.

Proof. For $t_o < t_i$, let $d = t_i - t_o$ and $m = (t_o + t_i)/2 - \pi$. The exact identity

$$f_a(t_o) - f_a(t_i) = 2a\cos(m)\sin(d/2) - d$$

applied to the endpoint pairs gives $G \le 0.4\varepsilon w_1$ for $w_1 > 0$. Let $x_a \in (0, \pi)$ solve $\sin x = x/a$, and $\beta = \arccos(1/a)$. The pairs $(-2, 0.8)$ and $(-0.8, 2)$ show that positive gap requires $0 < 1.4w_1 < x_a$ and, in the respective orientation, $|m_j| < \beta$ after a period shift, where $m_j = b_1 \mp 0.6w_1 - \pi$.

Taylor bounds imply $x_a^2 \le 6\varepsilon$ for $\varepsilon \le 0.6878$, and $\beta^2 \le \varepsilon(2+\varepsilon)$ follows from $\arctan z \le z$. To see the former, $\sin x/x \le 1 - (x^2/6)(1 - x^2/20)$ first gives $x_a^2 < 11.846\varepsilon$ using $x_a < \pi$, and substitution gives

$$x_a^2 \le \frac{6\varepsilon}{(1+\varepsilon)(1 - 0.5923\varepsilon)} \le 6\varepsilon.$$

Hence $p < \sqrt{6}/1.4$. Put $\mu = m_j/\sqrt{\varepsilon}$ and $c' = 1.96/6$. Multiplying the positive Taylor upper bounds on $\cos m_j$ and $\sin(1.4w_1)/(1.4w_1)$ gives

$$\mu^2/2 + c'p^2 < \Lambda(\varepsilon), \quad \Lambda = \frac{1}{(1+\varepsilon)\min(\kappa_1, \kappa_3)},$$

where $\kappa_1 = 1 - \varepsilon(2+\varepsilon)/12$ and $\kappa_3 = (1 - 0.3\varepsilon)(1 - \varepsilon(2+\varepsilon)/2)$. Since $q = \mu \pm 0.6p$, Cauchy–Schwarz bounds $|q| < \sqrt{(2 + 0.36/c')\Lambda}$. For $\varepsilon \le 0.029$, $\Lambda \le 1.03935$, so $|q| < 1.7956$.

The cubic witness $(p, q) = (8/5, 6/5)$ has gap $27088/46875$, by exact endpoint evaluation and comparison with the critical points $\pm\sqrt{2}$. Taylor's remainder gives

$$|\phi_\varepsilon - h| \le \varepsilon \left| \frac{\sigma^3}{6} - \frac{(1+\varepsilon)\sigma^5}{120} \right| + \frac{(1+\varepsilon)\varepsilon^2|\sigma|^7}{5040}.$$

On that witness, $|\sigma| \le 4.4$ and the first polynomial is bounded by 2.7713. Since a gap is 2-Lipschitz in the profile,

$$K(\varepsilon) \ge K_{\text{low}}(\varepsilon) = \frac{27088}{46875} - 2\varepsilon \left[2.7713 + (1+\varepsilon)\varepsilon \frac{4.4^7}{5040} \right].$$

This decreasing bound is 0.4062 at $\varepsilon = 0.029$. Positivity and the input-weight bound restrict the supremum to a compact domain modulo bias periods; continuity gives attainment. At a maximizer, $0.4p \ge K(\varepsilon) \ge 0.4062$, proving the lower bound on p.

Both the finite and limiting maximizers lie in $[1, 1.7497] \times [-1.7956, 1.7956]$, where $|\sigma| \le 5.295$. For the limit, use the cubic pair identity to obtain $G_0 \le 0.4|p|$ and $(q \mp 0.6p)^2 + (2.8p)^2/12 < 2$ at positive gap. The same uniform remainder and the 2-Lipschitz gap bound give $|K(\varepsilon) - K| \le 23.3\varepsilon$. □

The proof localizes the gap maximizer; it does not identify the tied corner. That identification needs the certified uniqueness of the cubic maximum, independently checked in ball arithmetic, followed by the parametric corner certificate. At $a = 1.02$–1.25, the localization lemmas combined with the certified gap lower bounds also place the maximizer in $\mathcal{C}$. The argument does not prove H1 for $a \ge 1.30$: its bias bound there exceeds 2. It concerns the continuous windows and the endpoint-containing population data, not arbitrary random samples or the finite-a loss minimizer.

F.2 First-order correction and its registered test

Writing $K(\varepsilon) = G^\star(1+\varepsilon)/\varepsilon^{3/2}$ and $A_\varepsilon = A^* + A'(0)\varepsilon + O(\varepsilon^2)$ gives

$$R_\varepsilon = \frac{A_\varepsilon K(\varepsilon)}{2}, \qquad c_1 = \frac{A'(0)}{A^*} + k_1, \qquad K(\varepsilon) = K(1 + k_1\varepsilon + O(\varepsilon^2)).$$

Differentiating the joint stationarity-and-zero-gap equations gives $A'(0)/A^* \in [0.6621547, 0.6621550]$. The gap maximum is a tied corner: the inner extrema tie at $x = \pm 0.8$, and the outer extrema at $x = -2.0, -1.2$. Differentiating its tie equations gives $k_1 = -0.37692472$ at the reported precision. Combining the enclosed terms gives $c_1 \in [0.2852300, 0.2852303]$.

A parametric Krawczyk certificate preserves the corner and its active set for the analytic continuation $|\varepsilon| \le 0.05$. The convex-hull weights proving a sharp local maximum are at least 0.00434. This makes the corner value analytic on both sides of zero. Identification with the global gap maximum uses separate global searches; active-set persistence alone proves a local statement. Lemma 2 identifies the searched-box maximum with the unrestricted gap maximum, enclosed in $[0.5794558833, 0.5794559217]$. Free searches at positive small ε support the same corner, without fixing its active set.

The first calculation omitted the switch shift and was rejected by the certified finite-a refit. The complete calculation was made after that refit had been seen. It has no fitted input, but was not blind to those data. The independent small-ε test was registered before computation at $a = 1.01$–1.04.

That registered test is *inconclusive*: the feasible set $[0.225, 0.330]$ has width 0.105, above the pre-set 0.1 limit. The predicted 0.285 lies inside. The set excludes 0, 0.49 and the two components alone, -0.377 and 0.662. The component k_1 test passed with feasible width 0.0013; the switch-shift

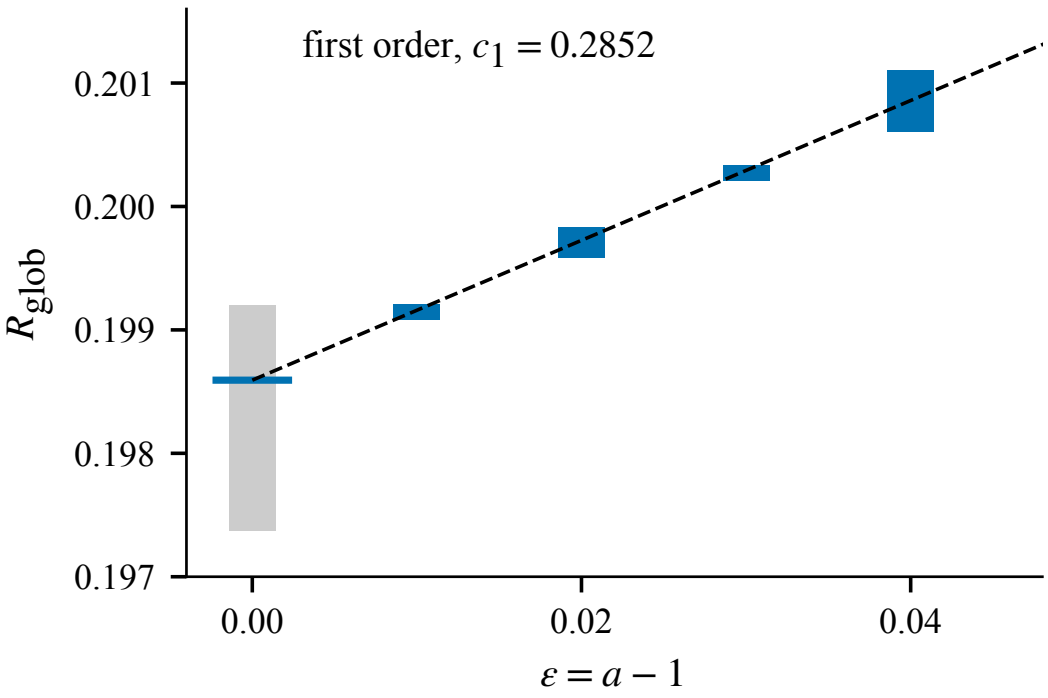


Figure 8: At small ε, certified thresholds follow the first-order line. This detail is omitted from Figure 3: four certified brackets at $a = 1.01$–1.04 and the limit, with no training runs. Boxes show search-certified deterministic enclosures to scale; independent finite-a placement checks cover $a = 1.30$–1.60 and the later $a = 1.08$–1.12 follow-up, not these plotted small-ε points. The coefficient is $c_1 \in [0.2852300, 0.2852303]$ and its first-order range is certified for $|\varepsilon| \leq 0.05$. The sharp, not fully independently checked value $[0.1985926, 0.1985927]$ lies inside the independently verified bracket $[0.19738, 0.19920]$; both use the global cubic-gap normalization established by Lemma 2. The registered test is inconclusive: feasible $c_1 \in [0.225, 0.330]$, width $0.105 > 0.1$.

component was inconclusive. A *supplementary* branch-root measurement gives $[0.2846, 0.2859]$ and carries no registered verdict.

For $\varepsilon = 0.3$–0.6, no single first-order term fits all certified thresholds. A two-term fit admits $c_1 \in [0.243, 0.321]$ and a negative quadratic term. A straight line through those finite-a thresholds extrapolates to 0.2002–0.2035, outside the certified limit. Exploratory tracking finds the same gap-maximizing active set at 79 values through $\varepsilon = 2.0$; a change that reverses between grid values would be missed. The observed concavity occurs while the corner persists and reflects higher-order terms, including the product of the two first-order corrections. It does not extend the certified local expansion to that whole range.

F.3 Registered follow-up of the first-order coefficient

The original test stays inconclusive. A follow-up registered after that result adds certified thresholds using the unchanged bracket procedure and feasible-set estimator. The decisive added points lie at $\varepsilon = 0.08$–0.12, within the registered $O(\varepsilon^3)$ allowance $|c_3| \leq 1$. This is not confirmation restricted to $\varepsilon \leq 0.04$.

All added thresholds certify. The estimator combines them with the original frozen brackets and gives $C = [0.26531, 0.30687]$, of width 0.0416. It contains the derived interval $[0.2852300, 0.2852303]$ and excludes 0, 0.49, -0.377 and 0.662. The registered follow-up passes. The original feasible width remains 0.105, above its validity limit. The supplementary branch-root interval $[0.2846, 0.2859]$ carries no verdict. All ten new endpoint certificates were independently

Table 7: New first-order follow-up certificates. Brackets are deterministic search-certified enclosures, not uncertainty intervals.

a	Output-scale bracket	$\widehat{G}(a)$	R_{glob} interval
1.08	$[31.837918, 31.91751275]$	$[0.01273485, 0.01273485]$	$[0.202725, 0.203232]$
1.09	$[26.865283375, 26.8819595]$	$[0.01514250, 0.01514251]$	$[0.203404, 0.203530]$
1.10	$[23.08107775, 23.0954585]$	$[0.01767344, 0.01767345]$	$[0.203961, 0.204088]$
1.11	$[20.118496, 20.143426]$	$[0.02031933, 0.02031935]$	$[0.204397, 0.204651]$
1.12	$[17.7671015, 17.778150688]$	$[0.02307296, 0.02307297]$	$[0.204970, 0.205097]$

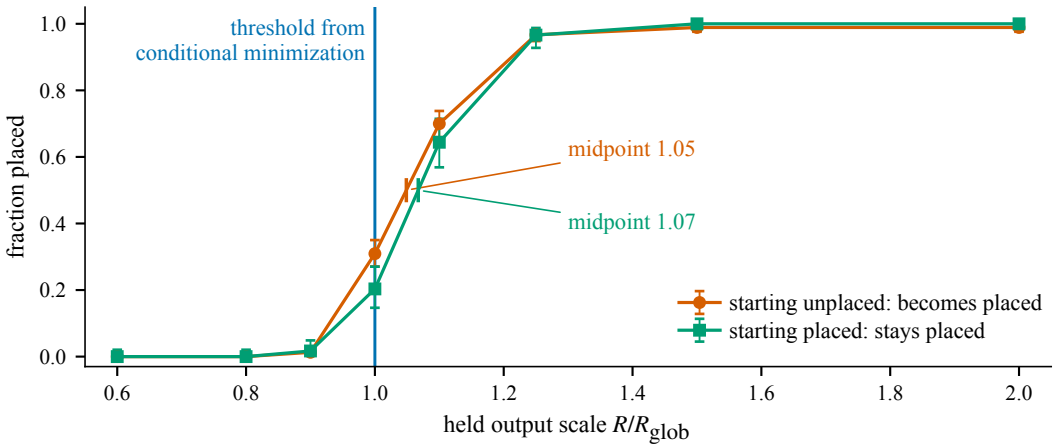


Figure 9: The fixed-scale placement and retention curves cross their midpoints just above the conditional threshold. These intervention curves supplement the unconstrained crossings in Figure 3. At $a = 1.30$, 543 pre-placement checkpoints per held level are read after 4,000 steps, while 177 already-placed runs per level are read after 12,000 steps. Adam's moments are preserved. Held levels R/R_{glob} are $0.6, 0.8, 0.9, 1, 1.1, 1.25, 1.5, 2$, using the certified bracket midpoint. Curve midpoints are 1.049 and 1.067, rounded to 1.05 and 1.07 in the plot. Intervals are Clopper–Pearson 95%; Figure 11 supplies the reset and longer-horizon curves.

checked in ball arithmetic and passed. The rescaled geometry enclosures remain search-certified only. None was excluded under the registered checker rule.

The bracket procedure stops at its first unresolved midpoint. Near the switch the competing minima meet tangentially; the loss tolerance prevents arbitrarily fine relative scale resolution. This is why the follow-up used larger positive epsilon. It does not measure the coefficient to the precision of its analytic derivation.

G Fixed-scale replays and mirror branches

Block 4 uses 543 pre-placement checkpoints from 180 fresh base-window runs at $a = 1.30$. Positive joint rescaling of (w_2, b_2) preserves every decision in all 8,688 replays. Output magnitude is then fixed while the other parameters train for 4,000 Adam steps, with moments preserved or reset. Block 5 starts at first placement in 177 runs and holds scale for 12,000 steps. Retention means positive gap at every 25-step check. Decisions are preserved in all 2,832 retention replays.

D1 passes for preserved and reset moments; D2 fails because placement does not fall at twice the threshold. The location and monotonicity predictions for Block 5 both pass. Its midpoint is 1.067 with preserved moments and 1.068 with reset moments. The 10–90% width is 0.27, reflecting heterogeneity across runs, not a measured adjustment timescale.

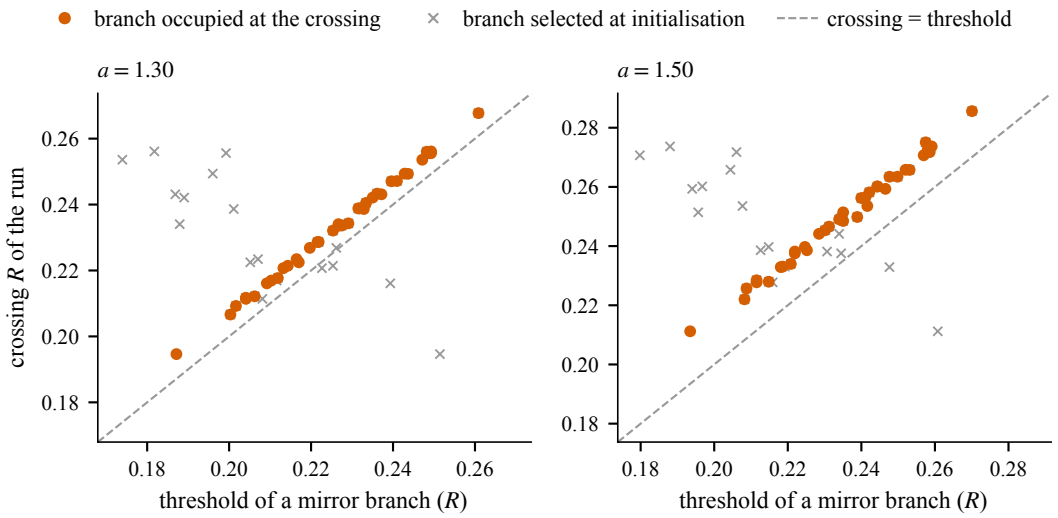


Figure 10: Post hoc, crossings track the threshold of the occupied mirror branch more closely than that of the branch selected at initialization. At budget 32,000, with every-step detection, each activation value has 38 crossing runs: 20 on branch + and 18 on branch −. Occupied-branch Spearman correlations are $0.997\ [0.9872, 0.9996]$ at $a = 1.30$ and $0.995\ [0.9801, 0.9983]$ at $a = 1.50$; brackets are bootstrap 95% intervals, with none drawn in the plot. Initial-branch correlations are 0.13 and 0.12; the own-global correlation is 0.88 at both values. Median lags $\log(R/T)$ are 0.030 and 0.063.

Table 8: Fixed-scale placement and retention in the base window, $a = 1.30$, under Adam with moments preserved. The reset variant has the same placement outcomes and the same registered retention verdicts.

R/R_{glob}	0.6	0.8	0.9	1.0	1.1	1.25	1.5	2.0
Placed, $n = 543$	.000	.000	.013	.309	.700	.967	.989	.989
Retained, $n = 177$	.000	.000	.017	.203	.644	.966	1.000	1.000

Horizon and own-threshold tests. At horizons 4,000, 16,000 and 64,000, placement midpoints are 1.0458, 1.0450 and 1.0454. Q1 and Q2 fail. Placement below the population threshold persists: at 64,000 steps, 8 of 543 replays place at $0.9R_{\text{glob}}$, and 9.4% place at $0.95R_{\text{glob}}$.

The own-sample rule agrees with 89.4% of 2,715 replays, versus 76.0% for the population rule, but misses its registered 95% criterion (S1). The midpoint differs from the replay-weighted median own threshold by 0.0095, inside the registered tolerance ± 0.02 (S2 passes). With placement checked every step, the own-sample excess accounts for the registered half of the free-training offset at $a = 1.30$, not at $a = 1.50$ (S3 passes and fails, respectively).

Table 9: Post hoc endpoint conditioning of the 64,000-step replays. The endpoint branch is known after the replay. These descriptive agreements explain the registered S1 failure; they do not replace it.

Endpoint	Replays	Own-global rule agreement	Disagreements
Preferred branch	2,097	.993	15
Other mirror branch	618	.558	273

All 41 below-own-threshold placements at levels 0.9 and 0.95 occupy the other mirror. Of the 15 preferred-branch disagreements, eight remain on the degenerate constant-predictor plateau; four are within 0.35% of their threshold and have not settled; three are below threshold-bracket resolution. No third basin is found. Excluding plateau endpoints leaves 2,078 preferred-branch replays with agreement 0.9966. With the threshold of the branch occupied at the *starting checkpoint*, agreement

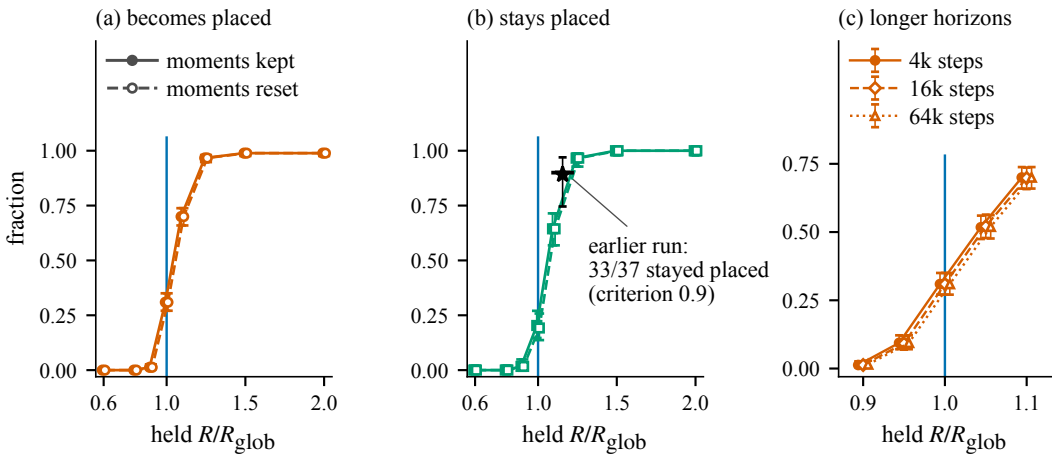


Figure 11: Resetting Adam's moments leaves the transitions in place, and longer horizons leave the placement midpoint unchanged. These checks are omitted from Figure 9. At $a = 1.30$, (a) uses 543 pre-placement checkpoints per level, read after 4,000 steps; preserved and reset curves coincide, with midpoints 1.049 and 1.049. Panel (b) uses 177 placed runs per level, read after 12,000 steps; midpoints are 1.067 and 1.068. The earlier high hold was 1.15 times an earlier threshold estimate, or $1.156R_{\text{glob}}$: 33 of 37 retained placement, failing the registered criterion of at least 0.9. Panel (c) uses 543 preserved-state replays per level on $R/R_{\text{glob}} = 0.90$–1.10; midpoints are 1.0458, 1.0450 and 1.0454 at 4,000, 16,000 and 64,000 steps. Both registered horizon predictions failed. All intervals are Clopper–Pearson 95%.

is 0.994; that branch is preserved in every replay. This is prediction from an observed starting state. Conditioning free training on its eventual crossing branch is instead an endpoint explanation.

The two mirror thresholds differ by median 0.106 times the global threshold, interquartile range 0.042–0.183, at $n = 400$. Their gap exceeds the distance from held scale to the own-global threshold for 281 of S1's 288 disagreements. An equal-weight mixture of branch thresholds underpredicts median free-training crossing R by 11–16%. The dynamics favor the lower-loss branch before commitment. The branch chosen at initialization matches the eventual branch in only 51% of runs; it is not a useful advance predictor.

Amended validity checks. The original replay check tried to freeze w_2 by zeroing its gradient inside Adam. Stored momentum still moved it. Before outcomes were computed, the reference was corrected to reset w_2 after every step; the replay then reproduced exactly. No outcome criterion changed. The horizon extension's reproduction gate also fired because parsed stored values differed at floating-point precision. Exact parsing restored bit-for-bit agreement in all 3,258 rows before scoring. The registration census records both amendments separately from outcome predictions.

H Sample size and interventions on the lag

The registered size test checks placement every step and uses 50 fresh seeds per cell at $a = 1.30, 1.50$ and $n = 400, 1{,}600, 6{,}400$. N1 passes at both activation values. N2 fails at 1.30 because the interval for the decrease includes zero, and passes at 1.50. N3 fails at both. The finite-sample threshold shift shrinks toward zero, to about 1% at the largest sample size. The free-training offset remains about 4% at $a = 1.30$ and 7.5% at $a = 1.50$. The mirror asymmetry also shrinks with sample size, whereas conditioning on the crossing branch leaves the lag essentially unchanged.

Changing output-weight growth. Both learning-rate tests check placement every step in every arm, at budget $32{,}000/\varphi$. The first registered test multiplies the output-weight learning rate by $\varphi \in \{0.25, 0.5, 1, 2\}$ from initialization on the size test's 50 samples per activation value at $n = 6{,}400$. Its primary L1 passes at both values; proportionality test L2 fails at both. Slowing the weight

Table 10: Sample-size test, 50 seeds per cell, with placement checked every step. Excess and offset are relative to the population threshold; brackets are the reported 95% intervals.

a	n	Own-threshold excess	Free-training offset
1.30	400	4.2% [1.8,7.4]	7.1% [3.7,9.8]
1.30	1,600	3.9% [1.5,5.1]	5.6% [3.2,7.3]
1.30	6,400	1.2% [0.5,2.2]	4.2% [3.1,5.2]
1.50	400	4.4% [1.4,7.3]	10.6% [8.3,13.7]
1.50	1,600	3.6% [1.2,5.1]	8.5% [6.9,11.0]
1.50	6,400	1.1% [0.2,2.1]	7.5% [6.4,8.3]

also lengthens time on the constant-predictor plateau, so the first test does not isolate the phase responsible for the effect.

The deconfounded primary intervention starts at $R \geq 0.7$ times the own threshold, after plateau exit. All arms therefore share the same preceding trajectory. The unchanged-rate continuation reproduces the original runs bit for bit in all 192 continuations. There are 48 eligible crossings per activation value. The ordering with rate holds, but quarter-rate reductions are only 11% and 14%, below the registered half-reduction criterion. L1′ and L2′ fail at both activation values. There is no plateau

Table 11: Lag after the late learning-rate intervention, 48 eligible runs per activation value, with placement checked every step in every arm. Each cell gives a median percentage and its reported 95% interval.

a	$\varphi = .25$	.5	1	2
1.30	2.78 [2.29,2.93]	3.02 [2.77,3.08]	3.11 [2.98,3.17]	3.16 [3.07,3.20]
1.50	5.63 [4.46,6.10]	6.29 [5.53,6.66]	6.56 [6.21,6.79]	6.67 [6.55,6.85]

re-entry in the primary late-switch arms and mirror share is unchanged. A secondary intervention beginning immediately after plateau exit reduces the lag to 0.71% and 1.64%. It is reported without a registered verdict. This comparison locates most of the growth-rate dependence earlier than the final approach; the post hoc tracking calculation is given in Appendix V.

Post hoc correlations and registered interventions. Block 4a is post hoc. At 0.7 of the own threshold, among 48 runs per activation value with crossings checked every step, the final lag correlates with displacement from the occupied branch's minimizer: Spearman $\rho = 0.68$ at $a = 1.30$ and 0.67 at $a = 1.50$, with permutation $p \approx 10^{-4}$. Correlations with Adam's $\sqrt{\widehat{v}(w_2)}$ are -0.47 and -0.48.

Block 4b intervenes at that point. The original checks reported: the control is bit-identical to the unchanged-rate continuation, the branch is preserved, moved parameters lie on the minimizer, no run is excluded and all 48 runs cross in each arm. Neither moving to the branch minimizer nor resetting Adam's moments removes half the lag, so both activation values return the registered competing outcome. The combined arm has no criterion.

Table 12: Block 4b as run, before the corrected optimizer-state rerun; placement checked every step, 48 crossings per arm and activation value. Entries are median lags against the own threshold. Brackets are 95% intervals for control minus arm, in percentage points; every interval includes zero. The combined arm is reported without a criterion.

a	Control	Move to branch	Reset moments	Both
1.30	3.105%	3.079% $[-0.05, +0.05]$	3.148% $[-0.16, +0.02]$	3.122% $[-0.14, +0.04]$
1.50	6.559%	6.547% $[-0.10, +0.17]$	6.722% $[-0.50, +0.02]$	6.601% $[-0.37, +0.16]$

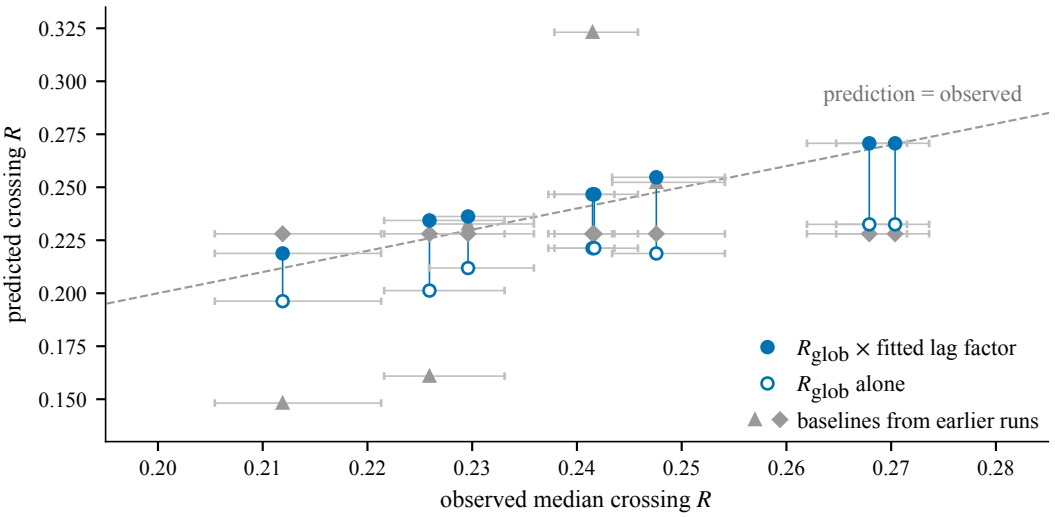


Figure 12: Across the held-out settings, the calibrated conditional threshold predicts median crossings better in aggregate than either baseline. Four windows at two activation values give eight settings, each with 86–88 crossings of 90 runs. Crossings and calibration use 50-step checks; horizontal bars are bootstrap 95% intervals for observed medians. The base-window fitted factors are $\lambda(1.30) = 1.115$ and $\lambda(1.50) = 1.164$. Triangles transfer the base crossing $|w_2|$; diamonds use the earlier pooled crossing $R = 0.2280$. Interpolating both calibration and observations preserves the sign of every registered comparison and excludes zero (Appendix I.2).

Corrected state initialization. A shared-state bug affected two arms of one block. A corrected rerun leaves its registered verdicts unchanged, and an audit of every other multi-arm experiment found no contamination. The move-to-branch arm inherited the control continuation's final Adam state instead of the checkpoint state; the move-and-reset arm inherited the reset continuation's final state instead of zero moments. Control and reset were unaffected. Fresh deep copies of the checkpoints were verified bit for bit before rerunning both affected arms. Corrected move-to-branch median lags are 0.03066 and 0.06504 at $a = 1.30, 1.50$, versus 0.03079 and 0.06547 as run. Corrected move-and-reset lags are 0.03130 and 0.06542, versus 0.03122 and 0.06601. Changes are under 1%; neither registered verdict changes. Both inherited-displacement and optimizer-memory predictions still fail. Table values above disclose the original arms; the corrected values here govern their interpretation.

Post hoc, the lag correlates with the run's displacement from its branch minimizer at the switch point; a registered intervention that removes the displacement does not remove the lag, so the correlation is not causal there. Neither inherited displacement nor optimizer memory alone accounts for it. The post hoc linear-response calculation in Appendix V accounts for these intervention arms without changing their registered verdicts.

I Prospective designs and complete setting inventory

Geometry-level test. The four windows are selected at the 0.1, 0.35, 0.65 and 0.9 quantiles of limiting geometric constants in a family of 88 windows, excluding every previously trained window. The activation values $1.30, 1.50$ match the calibration protocol. One timing-only H35 run, seed 999999, preceded registration. It is excluded from scoring, and no prediction depends on it. Predictions for the scored cohorts were frozen before their training. Thus the test is prospective for its scored runs, without a claim that no computation of any kind occurred on any selected geometry.

The calibration factors $\lambda(1.30) = 1.11487$ and $\lambda(1.50) = 1.1644$ are fitted on the base window only, with placement checked every 50 steps in both calibration and prospective runs. U and C use the same certified-bracket midpoint convention; bracket resolution contributes at most $\pm 0.6\%$. For the registered 50-step checks, the primary score is absolute log error of the setting's median placement-crossing R. The primary paired bootstrap resamples settings. The secondary window-clustered cal-

Table 13: Every geometry-test setting (Block 3), with placement checked every 50 steps and 90 runs per setting. Entries after observed crossing R are absolute log prediction errors. C improves the aggregate comparison, but loses to a baseline in three settings.

Window	a	Crossers	Observed R	U	C	B1	B2
H10	1.30	88	.2119	.077	.032	.358	.073
H10	1.50	88	.2259	.116	.037	.339	.009
H35	1.30	86	.2296	.080	.028	.013	.007
H35	1.50	87	.2476	.124	.028	.019	.082
H65	1.30	87	.2415	.088	.021	.291	.057
H65	1.50	87	.2679	.142	.010	.267	.161
H90	1.30	88	.2417	.088	.021	.492	.058
H90	1.50	88	.2704	.151	.001	.458	.170

culation keeps the two activation values together: C minus B1 has interval $[-0.414, -0.069]$ and C minus B2 has interval $[-0.098, -0.012]$. Leave-one-window-out comparisons retain the advantage. C overpredicts all settings; no held-out recalibration is performed.

With placement checked every 50 steps, the U-offset range $[0.080, 0.182]$ contains seven settings. H10 at $a = 1.30$ gives 0.0767. Its partial census verdict was assigned post hoc because the registration specifies per-setting inside/outside outcomes without an aggregate verdict. Under a joint interpretation the prediction fails. The primary comparison is unaffected.

Own-sample test. The second design uses V10–V80 at $a = 1.30, 1.50$, with 60 runs per setting and 200 points per run. Thresholds are computed from each training set before unconstrained training and frozen by SHA-256. Certified endpoint checks on 48 runs found no contradictions. P1–P4 pass under their registered rules at both activation values with placement checked every 50 steps. P2b still passes after interpolation, but C's advantage at $a = 1.50$ does not: the registered comparison is $+0.023\,[+0.009, +0.036]$ and the interpolated comparison is $-0.003\,[-0.013, +0.008]$. Tables below report all check-based per-setting errors. Appendix I.2 gives the sensitivity analysis without changing registered verdicts.

Table 14: Own-sample prospective errors at $a = 1.30$, with placement checked every 50 steps. Entries are mean absolute log errors and bootstrap 95% intervals, using 10,000 resamples with seed 0. Per-setting run-bootstrap intervals are descriptive; the all-window row uses the registered window bootstrap. Every run has a defined S-early predictor.

Window	Crossed	U_{own}	U	C	C_{own}	S-early
V10	57/60	0.046 [0.034, 0.061]	0.121 [0.100, 0.143]	0.086 [0.070, 0.104]	0.022 [0.008, 0.039]	0.075 [0.050, 0.104]
V20	57/60	0.053 [0.039, 0.070]	0.115 [0.095, 0.136]	0.082 [0.065, 0.100]	0.028 [0.012, 0.048]	0.072 [0.047, 0.104]
V30	57/60	0.052 [0.039, 0.067]	0.109 [0.091, 0.128]	0.073 [0.058, 0.090]	0.027 [0.012, 0.045]	0.070 [0.047, 0.098]
V40	57/60	0.053 [0.042, 0.067]	0.100 [0.083, 0.117]	0.065 [0.052, 0.079]	0.026 [0.013, 0.043]	0.070 [0.050, 0.095]
V50	58/60	0.056 [0.047, 0.068]	0.166 [0.143, 0.190]	0.091 [0.074, 0.111]	0.029 [0.017, 0.045]	0.070 [0.054, 0.090]
V60	58/60	0.052 [0.044, 0.063]	0.146 [0.126, 0.167]	0.075 [0.060, 0.091]	0.025 [0.015, 0.039]	0.074 [0.055, 0.098]
V70	58/60	0.056 [0.050, 0.064]	0.116 [0.100, 0.131]	0.053 [0.043, 0.065]	0.028 [0.020, 0.040]	0.079 [0.062, 0.097]
V80	58/60	0.053 [0.046, 0.061]	0.104 [0.092, 0.115]	0.042 [0.033, 0.053]	0.026 [0.017, 0.037]	0.070 [0.054, 0.089]
All 8	460/480	0.052 [0.050, 0.054]	0.122 [0.109, 0.138]	0.071 [0.060, 0.081]	0.026 [0.025, 0.028]	0.072 [0.070, 0.075]

The registered geometry-level calibration is $C = \lambda(a) R_{glob}$, where λ is the base window's median crossing R divided by its certified R_{glob}, with crossings checked every 50 steps. It was frozen before

Table 15: Own-sample prospective errors at $a = 1.50$, with placement checked every 50 steps. Entries are mean absolute log errors and bootstrap 95% intervals, using 10,000 resamples with seed 0. Per-setting run-bootstrap intervals are descriptive; the all-window row uses the registered window bootstrap. Every run has a defined S-early predictor.

Window	Crossed	U_{own}	U	C	C_{own}	S-early
V10	57/60	0.079 [0.066, 0.096]	0.143 [0.120, 0.167]	0.088 [0.070, 0.107]	0.033 [0.016, 0.054]	0.102 [0.079, 0.128]
V20	57/60	0.085 [0.074, 0.099]	0.142 [0.121, 0.164]	0.079 [0.061, 0.097]	0.033 [0.017, 0.053]	0.098 [0.080, 0.122]
V30	57/60	0.087 [0.078, 0.097]	0.139 [0.119, 0.159]	0.069 [0.054, 0.084]	0.033 [0.020, 0.050]	0.113 [0.090, 0.141]
V40	57/60	0.092 [0.084, 0.101]	0.143 [0.127, 0.160]	0.058 [0.047, 0.070]	0.037 [0.025, 0.051]	0.118 [0.097, 0.142]
V50	58/60	0.103 [0.095, 0.112]	0.219 [0.193, 0.246]	0.098 [0.080, 0.117]	0.046 [0.035, 0.060]	0.125 [0.108, 0.146]
V60	58/60	0.108 [0.099, 0.117]	0.204 [0.182, 0.227]	0.084 [0.070, 0.099]	0.051 [0.040, 0.063]	0.136 [0.117, 0.159]
V70	58/60	0.108 [0.099, 0.118]	0.170 [0.152, 0.187]	0.061 [0.050, 0.073]	0.051 [0.040, 0.063]	0.138 [0.121, 0.159]
V80	58/60	0.109 [0.098, 0.120]	0.167 [0.151, 0.183]	0.055 [0.046, 0.065]	0.052 [0.041, 0.065]	0.133 [0.115, 0.154]
All 8	460/480	0.096 [0.089, 0.104]	0.166 [0.148, 0.188]	0.074 [0.064, 0.084]	0.042 [0.036, 0.048]	0.120 [0.111, 0.130]

Table 16: Prospective geometries: class 0 is $I = [-i, i]$ and class 1 is $O = \pm[l, u]$. H windows belong to Block 3; V windows belong to the own-sample test. The ν_0 intervals are deterministic geometric enclosures.

Window	i	l	u	ν_0
H10	0.6	0.8	2.0	$[0.19101, 0.19102]$
H35	0.6	0.9	1.7	$[0.30072, 0.30074]$
H65	0.8	1.4	2.2	$[0.41889, 0.41891]$
H90	0.7	1.5	2.7	$[0.51326, 0.51329]$
V10	0.7	0.9	1.5	$[0.19030, 0.19031]$
V20	0.8	1.1	1.9	$[0.24159, 0.24161]$
V30	0.9	1.3	2.1	$[0.27897, 0.27899]$
V40	0.9	1.4	2.0	$[0.33443, 0.33445]$
V50	0.5	0.9	2.1	$[0.35854, 0.35856]$
V60	0.6	1.1	2.3	$[0.39845, 0.39847]$
V70	0.5	0.9	1.5	$[0.43300, 0.43304]$
V80	0.9	1.7	2.3	$[0.47250, 0.47254]$

Block 3. For the own-sample test,

$$C_{own} = \rho_{lag}(a) R_{own}, \qquad \rho_{lag}(a) = \exp\left(\text{median} \log \frac{R_{cross}}{R_{own}} \right),$$

with $\rho_{lag}(1.30) = 1.0311$ and $\rho_{lag}(1.50) = 1.066$. This fitted lag factor uses base-window Phase 2b seeds 0–39 that crossed within 12,000 steps, with placement checked every step; it was frozen before any computation on the new settings, whose crossings are checked every 50 steps.

With placement checked every 50 steps, the 20 non-crossers at each activation value are included at final R in a sensitivity calculation, which preserves the comparisons. This is a sensitivity analysis for censored runs, not an observed crossing for them. Per-setting Spearman correlations range from 0.76 to 0.86 at $a = 1.30$ and 0.57 to 0.82 at $a = 1.50$.

In the 50-step-check own-sample test, the prospective-in-time S-early rule uses the branch at first plateau exit. It matches the crossing branch in 87% and 88% of runs and precedes crossing by

median 2,250 and 1,000 steps. Its median errors 0.044 and 0.092 lie in the registered ranges. The fitted S-mix predictor beats U but not C; it has no criterion and receives no pass verdict.

I.1 Stronger baselines, assessed post hoc

Each baseline was fitted only on the earlier calibration settings. Pooling the lag factor over all five calibration windows gives mean absolute log error 0.0231, versus 0.0223 for the registered predictor; the paired difference is -0.0007 $[-0.0088, +0.0074]$, so they are not separated. Taking the median of the windows' factors gives 0.0274 and is also not separated. Baselines that do not use the conditional threshold do 4–5 times worse than the registered predictor. A variant that pools the lag factor over all calibration windows matches it (post hoc). These results use the earlier 50-step detections; Appendix I.2 gives the fully interpolated sensitivity analysis for the registered comparisons.

Table 17: Post hoc baselines on the held-out geometry population, with 50-step crossing checks. Errors are mean absolute log errors. The last column is the paired-bootstrap 95% interval for registered predictor minus baseline.

Predictor	Error	Difference interval
Registered conditional threshold, base factor	0.0223	–
Conditional threshold, pooled factor	0.0231	$[-0.0088, +0.0074]$
Conditional threshold, median factor	0.0274	$[-0.0120, +0.0019]$
Pooled crossing R	0.0773	$[-0.1027, -0.0105]$
Regression on $\log \widehat{G}$ at each a	0.1047	$[-0.0926, -0.0722]$
Regression on $\log \nu_0$ and a	0.1099	$[-0.0993, -0.0746]$
Base-window crossing weight	0.2796	$[-0.3704, -0.1312]$

I.2 Crossing detection and sensitivity

Phase 2b (S3), the sample-size test, both learning-rate tests and Block 4b check placement every step. The learning-rate interventions retain that check interval at every φ, although their budgets are $32{,}000/\varphi$. Block 3, the own-sample prospective test and the base calibration runs behind λ, B1 and B2 check every 50 steps.

For every-step detection, the recorded crossing's upward bias is at most one step's growth: median 0.05–0.15% of the threshold, at most 0.44%. Interpolation changes the own-threshold lag by at most 0.09 percentage points. S3 remains 3.1% and 6.6% at $a = 1.30, 1.50$; the size-test lags move from 3.2% to 3.1% and from 6.7% to 6.6%. Check spacing does not explain this every-step lag.

For the 50-step experiments, the post hoc sensitivity analysis linearly interpolates the gap between the last negative and first positive check. Table 18 separates that analysis from registered check-based statistics. The lag remains after interpolation: 2.9% and 6.2% against each run's own threshold, and 7.3% and 11.5% against the population threshold in Block 3. In the own-sample test, detection adds about 1.2 and 2.6 percentage points at the two activation values. P2b passes un-

Table 18: Post hoc interpolation sensitivity for experiments checked every 50 steps. These are descriptive point statistics; registered verdicts retain their original detection definition.

Quantity	Check-based	Interpolated
Own-sample median $R_{\text{cross}}/U_{\text{own}} - 1$, $a = 1.30$, $n = 460$	4.1%	2.9%
Own-sample median $R_{\text{cross}}/U_{\text{own}} - 1$, $a = 1.50$, $n = 460$	8.8%	6.2%
Block 3 median setting excess over U, $a = 1.30$	8.8%	7.3%
Block 3 median setting excess over U, $a = 1.50$	14.2%	11.5%
Base calibration $\lambda(1.30)$	1.1149	1.0979
Base calibration $\lambda(1.50)$	1.1644	1.1298
B2 pooled crossing R over five base-calibration windows	0.2280	0.2260

Table 19: Own-sample comparisons: 50-step checks versus post hoc interpolation, 460 crossings per activation value. Brackets for P1–P3 are 95% intervals; P4 brackets are registered acceptance ranges, not confidence intervals. Each comparison passes under both definitions.

Criterion	Check-based	Interpolated
P1, a = 1.30	$-0.0696[-0.0852, -0.0565]$	$-0.0712[-0.0863, -0.0585]$
P2a, a = 1.30	$-0.0136[-0.0256, -0.0016]$	$-0.0302[-0.0372, -0.0226]$
P3, a = 1.30	$-0.0448[-0.0553, -0.0334]$	$-0.0523[-0.0593, -0.0447]$
P4, a = 1.30	$0.0419[0.0010, 0.0610]$	$0.0293[0.0010, 0.0610]$
P1, a = 1.50	$-0.0695[-0.0864, -0.0562]$	$-0.0698[-0.0861, -0.0568]$
P2b, a = 1.50	$0.0226[0.0091, 0.0363]$	$-0.0025[-0.0131, 0.0084]$
P3, a = 1.50	$-0.0318[-0.0436, -0.0196]$	$-0.0468[-0.0530, -0.0397]$
P4, a = 1.50	$0.0867[0.0340, 0.0940]$	$0.0615[0.0340, 0.0940]$

der both definitions. The stronger claim that C beats $\mathrm{U}_{\mathrm{own}}$ at $a = 1.50$ depends on detection: the positive check-based interval becomes an interval containing zero under interpolation.

For Block 3, the sensitivity analysis is fully interpolated: both the base calibration behind λ, B1 and B2 and the held-out observations use interpolated crossings. It preserves the sign of every registered comparison and excludes zero. Table 20 also reports the observations-only calculation, which mixes detection rules and is not the sensitivity analysis used here.

Table 20: Block 3 sensitivity: upper endpoints of the 95% intervals for differences in absolute log error. Negative values favor C. The registered analysis uses 50-step checks; the fully interpolated analysis uses interpolation for both calibration and observations. The last column mixes the two detection rules.

Comparison	Check-based	Fully interpolated	Observations only
C − B1	-0.1312	-0.1349	-0.1270
C − B2	-0.0105	-0.0169	$+0.0053$
C − U	-0.0630	-0.0544	-0.0363

J Interventions on output scale

For a positive factor k, replacing (w_2, b_2) by (kw_2, kb_2) multiplies every logit by k and preserves every decision at the intervention. We apply this joint rescaling before either holding the output-weight magnitude fixed or releasing it. Subsequent placement changes therefore follow the altered dynamics rather than an instantaneous change in classifications.

Reducing and holding R below the placement threshold removes positive gap by the first checkpoint, within 50 steps in all 37 eligible runs. Holding it above retains placement in 33 of 37, short of the registered $9/10$ criterion. E-2 therefore fails, and Block E's primary result is negative under its registered rule. The Fisher contrast ($p = 1.2 \times 10^{-16}$) does not change that verdict. The high hold is 1.156 times the rebuilt certified placement threshold and lies below the solve threshold; no run solves, consistent with the conditional branch's positive gap and incorrect classifications at that scale. This observation concerns the measured branch and intervention, not all parameters at that scale.

Supplying scale early gives the opposite of our prediction: 9 of 40 runs place, versus 37 of 40 controls ($p = 1.3 \times 10^{-10}$). The registered diagnostic supports saturation for this released-scale arm. Gradients fall to 0.072% of control, and extending the budget from 12,000 to 60,000 steps raises placement from 2 of 15 to 7 of 15. The held-high cold-start arm is mixed under its registered rule: gradients also collapse, but only 3 of 15 runs place at the longer budget. Neither result supports the predicted trapping explanation.

A separate growth-rate intervention failed its validity gate: its nominal control differed from unconstrained training by 42.5 percentage points. Its dose-response is therefore not evidence for a growth-rate mechanism. The retention contrast remains a measurement, but the registered claim

Table 21: Earlier Adam interventions at $a = 1.30$; scale multipliers refer to the then-frozen threshold, with 40 seeds per arm and 12,000 steps. Retention is measured among the 37 runs that had achieved placement. The early-intervention arms use all 40 seeds.

Intervention	Placement outcome	Registered prediction
Hold at $0.85R_{\text{glob}}^{\text{old}}$ after placement	0/37 retained	Loss of placement
Hold at $1.15R_{\text{glob}}^{\text{old}}$ after placement	33/37 retained	Retention (failed)
Negligible rescaling after placement	37/37 retained	No change
Supply scale early, then release	9/40 placed	Earlier placement (failed)
Start high and hold scale	2/40 placed	Placement (failed)
Unconstrained control	37/40 placed	Reference

that the threshold governs placement stability is not supported. A large initial scale also does not guarantee discovery of correct placement.

K THE SMALL-SCALE CRITERION AND WIDTH TWO

For balanced logistic classification, let ϕ_θ be the hidden output at fixed hidden parameters and let $\Delta\mu(\theta) = \mathbb{E}[\phi_\theta \mid y = 1] - \mathbb{E}[\phi_\theta \mid y = 0]$, with expectations over the training sample. Profiling the output bias gives the expansion

$$\mathcal{L}^*(\theta; s) = \log 2 - \frac{s}{4}\Delta\mu(\theta) + \frac{s^2}{8}\operatorname{Var}(\phi_\theta) + O(s^4). \tag{9}$$

Lemma 4 (Explicit small-scale remainder)**.** *For balanced labels, let $X = u_\theta - \mathbb{E}u_\theta$, let $D(\theta) = \max u_\theta - \min u_\theta$ over the data, and let $B_\theta(s) = \log 2 - \frac{s}{4}\Delta\mu(\theta) + \frac{s^2}{8}\operatorname{Var}(u_\theta)$. Then*

$$B_\theta(s) - \frac{s^4 D(\theta)^4}{192} \;\le\; L^*(\theta; s) \;\le\; B_\theta(s).$$

Proof. Absorbing the mean into the bias β, the balanced logistic loss is $L^*(\theta; s) = \log 2 - \frac{s}{4}\Delta\mu(\theta) + \min_\beta \mathbb{E}\log\cosh\frac{sX+\beta}{2}$. For all real z, $z^2/2 - z^4/12 \le \log\cosh z \le z^2/2$. Taking $\beta = 0$ gives the upper bound. At the minimizing bias, $\mathbb{E}\tanh\frac{sX+\beta^*}{2} = 0$, so $\beta^* \in [-s\max X, -s\min X]$ and $|sX + \beta^*| \le sD(\theta)$. The lower inequality and $\mathbb{E}(sX + \beta^*)^2 = s^2\operatorname{Var}(u_\theta) + \beta^{*2}$ give the lower bound. □

Small-scale selection. Let θ_s minimize $L^*(\cdot; s)$ and let θ_0 be any maximizer of $\Delta\mu$, with maximum Δ^*. By Lemma 4, $B_{\theta_s}(s) - s^4 D(\theta_s)^4/192 \le L^*(\theta_s; s) \le L^*(\theta_0; s) \le B_{\theta_0}(s)$, so

$$\frac{2}{s}\big(\Delta^* - \Delta\mu(\theta_s)\big) + \operatorname{Var}(u_{\theta_s}) - \operatorname{Var}(u_{\theta_0}) \le \frac{s^2 D(\theta_s)^4}{24}.$$

On the common compact set, $D(\theta_s)$ and the variances are bounded, so $\Delta\mu(\theta_s) \to \Delta^*$. By continuity every limit point $\bar\theta$ maximizes $\Delta\mu$, and since the first term is nonnegative, $\operatorname{Var}(u_{\bar\theta}) \le \operatorname{Var}(u_{\theta_0})$ for every maximizer θ_0.

Lemma 5 (Large-scale placement and the small-scale criterion)**.** *For the balanced finite sample, define the data gap*

$$G_n(\theta) = \min_{i:y_i=1} \phi_\theta(x_i) - \max_{i:y_i=0} \phi_\theta(x_i), \qquad \Gamma_n = \sup_\theta G_n(\theta).$$

Suppose $\Gamma_n > 0$ is attained at θ_Γ. Then

$$\mathcal{L}^*(\theta_\Gamma; s) \le \log(1 + e^{-s\Gamma_n/2}),$$
$$\mathcal{L}^*(\theta; s) \ge \frac{1}{n}\log(1 + e^{-sG_n(\theta)/2}).$$

Every attained conditional minimum is placed on the data for sufficiently large s. If the minimizers remain in a compact set as $s \to \infty$, their limit points maximize G_n.

Under the small-scale selection hypotheses above, if all variance-selected limit points have strictly negative G_n, conditional minimizers are unplaced at sufficiently small scales and placed at sufficiently large scales. If all selected limit points instead have positive continuous-window gap G, they are placed at sufficiently small scales, so there is no unplaced small-scale regime.

Proof. At θ_Γ, placing the bias at the midpoint of the data gap gives every point margin at least $s\Gamma_n/2$, proving the upper bound. At any other placement and bias, the extreme points from the two classes cannot both have margin exceeding $sG_n(\theta)/2$. One such point contributes at least $n^{-1}\log(1+e^{-sG_n(\theta)/2})$.

For $G_n \le 0$, this lower bound is at least $\log 2/n$, whereas the upper bound tends to zero. Thus large-scale minimizers must have positive data gap. For $0 < \delta < \Gamma_n$, the same comparison excludes $G_n \le \Gamma_n - \delta$ once $s > (2/\delta)\log(2n)$, proving the limit-point statement. At small scales, compactness and continuity preserve a strict gap sign shared by all selected limit points. The negative case therefore requires a placement change before the large-scale regime; the positive case excludes an unplaced regime extending to zero scale. □

The finite-sample bound concerns $G_n \ge G$, not the continuous-window gap itself. This endpoint argument establishes a placement change, but does not establish a unique crossing or settle a zero-gap small-scale limit. Continuous-window placement and the width-1 crossing are checked separately.

Width-2 parameterization and variance selection. At width 2, write

$$\phi_\theta(x) = \sum_{i=1}^{2} \widetilde{v}_i f_a(\alpha_i x + \beta_i), \qquad \|\widetilde{v}\|_1 = 1, \qquad N(x) = s\phi_\theta(x) + b, \quad s = \|w_2\|_1 > 0.$$

For the continuous symmetric windows used in the small-scale selection statement, define

$$D(\alpha) = \mathbb{E}_O \cos(\alpha x) - \mathbb{E}_I \cos(\alpha x), \qquad \alpha^* \in \arg\max_{\alpha>0} |D(\alpha)|.$$

Symmetry removes the linear contribution to $\Delta\mu$. For this width-two small-scale selection statement, expectations are over the continuous windows. Each active unit in a first-order maximizing configuration has $\alpha_i = \pm\alpha^*$ and its oscillatory part aligned with $D(\alpha^*)$. On this maximizing set,

$$\phi_\theta(x) = cx + \text{const} + a\,\text{sign}(D(\alpha^*))\cos(\alpha^* x), \qquad c = \sum_i \widetilde{v}_i\alpha_i.$$

Since $\text{Cov}(x, \cos(\alpha^* x)) = 0$,

$$\text{Var}(\phi_\theta) = c^2\,\text{Var}(x) + a^2\,\text{Var}(\cos(\alpha^* x)).$$

The variance-selected pair has $c = 0$, with $\widetilde{v}_1\alpha_1 = -\widetilde{v}_2\alpha_2$. A single unit cannot make this cancellation; its second-order loss excess is $(s^2/8)\alpha^{*2}\,\text{Var}(x)$.

Output-scale normalization and search domain. For an activation u, define the maximal continuous-window gap

$$\Gamma_2(u) = \sup_{\|\widetilde{v}\|_1=1,\,\theta} G_+\left(\sum_{i=1}^{2} \widetilde{v}_i u(\alpha_i x + \beta_i)\right), \qquad R_2 = \frac{\|w_2\|_1\widehat{\Gamma}_2}{2}, \quad s = \frac{2R_2}{\widehat{\Gamma}_2}.$$

Here $G_+(\phi) = \min_O \phi - \max_I \phi$ and $\widehat{\Gamma}_2$ is a validated numerical value, not a certificate of Γ_2. The normalization uses the lower values from the Nelder–Mead search: 1.1868728875 at $a = 1.30$ and 1.3694651983 at $a = 1.50$. The independent differential-evolution search differs by 3.3×10^{-6} and 7.2×10^{-7} relative, respectively. For tanh, $\Gamma_2 = 1$ and $R_2 = \|w_2\|_1/2$.

The conditional loss uses $\widetilde{v} = (t, \sigma(1-|t|))$, with $t \in [-1,1]$ and $\sigma \in \{-1,1\}$, and profiles the output bias. For f_a, starting values of each α_i are uniform on $[0,10]$ and each β_i on $[0,2\pi)$; t starts uniformly on $[-1,1]$ and σ is drawn at random. For tanh, the starting β_i range is $[-10,10]$. These boxes restrict initialization only: the hidden-parameter search is unconstrained, while t is clipped to $[-1,1]$. Results with α_i outside the starting box are flagged rather than excluded. The

retained cancelling pair has $\alpha \approx 1.79$. An independent CMA-ES search starts α_i, β_i in $[-10, 10]$ and normalizes a weight vector drawn from $[-1, 1]^2$, using a different parameterization.

The loss is evaluated on the 800-point population objective: 400 inner points on $[-0.8, 0.8]$ and 200 points on each outer interval. The validated direct check covers $R_2 \in \{0.001, 0.003, 0.01, 0.02\}$ at both $a = 1.30$ and $a = 1.50$. A separate threshold scan belonged to the width-2 design, whose registration step was never reached. At $a = 1.30$, it never started: its stop rule fired when the two searches for $\widehat{\Gamma}_2$ disagreed. At $a = 1.50$, it completed $R_2 = 0.02, 0.03, \ldots, 0.10$, nine of 99 planned points, before termination. Each point used 2,000 restarts and passed its audit; every minimizer was placed, with G_+ from 1.0262 to 1.0264. These scan points were not validated in the design's sense.

Placement criterion. Both the small-scale verdict and the direct checks use the continuous-window class gap G_+. Extrema over I and O are computed as enclosures, and a configuration is placed when the lower end of its gap enclosure is positive. The small-scale calculation also records the data-point gap G_n; the direct checks do not compute G_n.

Registered width-1 and width-2 results. The width-1 consistency check passes: the selected non-aliased class-mean-gap maximizer is unplaced at all six activation values, with gaps ranging from -3.349 to -3.662. At width 2 we find no placement threshold at any scale searched: the validated direct check at $R_2 \in \{0.001, 0.003, 0.01, 0.02\}$ at both activation values, and the unvalidated scan from $R_2 = 0.02$ to 0.10 at $a = 1.50$. The registered small-scale decision rule selects the placed cancelling pair, and all eight direct checks find that pair as the conditional minimizer. The pair cancels the linear part of the activation; its continuous-window and data-point gaps agree at $+0.8896$ for $a = 1.30$ and $+1.0265$ for $a = 1.50$.

The prediction rests on a second-order selection: at first order, unplaced single units and placed pairs tie. The variance selection was added to the registration before any maximizer was computed; it was absent from the original proposed rule. The width-2 outcome is a registered validated computation, not a certificate. Its registration specified a decision rule with two outcomes rather than predicting an outcome; the observed absence of a threshold is therefore listed separately from pass/fail predictions. W1 and W4 were not run. The direct check is complete at all eight scales. Every validation passed, and no finished restart improved on the retained cancelling pair. These conditional searches do not test training dynamics; the separate registered training tests are reported in Appendix N.

Unplaced local minima and the boundary infimum. The following comparisons describe the numerical searches. At these scales the unplaced region contains local minima: single units at α^* with an idle second unit ($G_+ \approx -3.3$ to -3.7), whose loss exceeds the placed cancelling pair's by the predicted single-unit second-order gap $(s^2/8)\alpha^{*2}\,\mathrm{Var}(x)$ to within 0.1–2%. The infimum over the unplaced region is lower: it lies on the placement boundary $G_+ = 0$, at a partially cancelling pair ($|c| \approx 0.43$–0.51). Descent from there reaches the placed cancelling pair, and the loss drop matches the second-order prediction $(s^2/8)c^2\,\mathrm{Var}(x)$ to within 0.2–3%. In every search the placed minimizer lies below every unplaced configuration, by at least 9.9×10^{-8} at $R_2 = 0.001$ (about $99\times$ the 10^{-9} tie tolerance).

Table 22: Direct small-scale conditional searches, 4,000 restarts per scale. Loss values are deterministic evaluated results, with tie tolerance 10^{-9}; no statistical intervals are attached. All eight checks are complete and validated; every retained minimizer is a placed cancelling pair.

a	R_2	Best placed loss	Best unplaced loss	Difference
1.30	0.001	0.6922871697904	0.6922872690039	9.92×10^{-8}
1.30	0.003	0.6905695256248	0.6905704168112	8.91×10^{-7}
1.30	0.01	0.6845827328908	0.6845925670975	9.83×10^{-6}
1.30	0.02	0.6760975221489	0.6761364675950	3.89×10^{-5}
1.50	0.001	0.6922871675820	0.6922872667961	9.92×10^{-8}
1.50	0.003	0.6905695190120	0.6905704102030	8.91×10^{-7}
1.50	0.01	0.6845827109908	0.6845925452478	9.83×10^{-6}
1.50	0.02	0.6760974787560	0.6761364244002	3.89×10^{-5}

Table 23: Unplaced boundary candidates and local minima from the same searches. The local-minimum loss difference is relative to the placed cancelling pair. All direct validations in Table 22 are complete.

a	R_2	Boundary G_+	Local-minimum loss difference	Local-minimum G_+
1.30	0.001	-2.0×10^{-10}	1.61×10^{-6}	-3.65
1.30	0.003	-1.2×10^{-11}	1.45×10^{-5}	-3.64
1.30	0.01	-2.0×10^{-12}	1.60×10^{-4}	-3.59
1.30	0.02	-1.6×10^{-13}	6.33×10^{-4}	-3.52
1.50	0.001	-1.7×10^{-10}	1.21×10^{-6}	-3.45
1.50	0.003	-1.9×10^{-12}	1.09×10^{-5}	-3.44
1.50	0.01	-4.4×10^{-13}	1.20×10^{-4}	-3.40
1.50	0.02	-2.9×10^{-13}	4.77×10^{-4}	-3.34

Table 24: Prospective own-sample and timescale tests with every-step detection. Brackets give paired 95% intervals for the mean absolute log-error difference, own threshold minus population threshold. Lag predictions use the frozen Adam fit; tolerances in row order are ± 0.0100, ± 0.0133, ± 0.0114 and ± 0.0205. All comparisons shown pass. For SGD, the separate correlation test fails when non-crossers are imputed.

Optimizer	a	Crossed	Own minus population	Predicted	Observed
SGD	1.30	30/40	$[-0.0738, -0.0172]$	0.0251	0.0295
SGD	1.50	30/40	$[-0.0727, -0.0148]$	0.0530	0.0643
Adam	1.45	80/80	$[-0.059, -0.022]$	0.0456	0.0532
Adam	1.60	80/80	$[-0.059, -0.022]$	0.0818	0.0809

The tanh test. For tanh the supremum of $\Delta\mu$ is not attained; the registered test returned "neither" because its boundary criterion failed at the largest box, although every other expectation held. Analytically, $\Delta\mu < 1 = \sup \Delta\mu$, so tanh lies outside the attainment hypothesis. Validation passes at every box size, the maxima rise toward one, and the first-order tie remains. The variance-selected symmetric pair is placed at every box, with $G_+ \to 1$, while the single units are unplaced. At $A = 40$, the 10^{-9} tie tolerance is of the order of $1 - \max \Delta\mu = 4.1 \times 10^{-9}$, causing the registered boundary criterion to fail. There was no re-registration.

L Prospective tests under SGD and at unseen activation values

L.1 SGD on samples with thresholds fixed in advance

The registered SGD test uses the same samples and initializations as the Adam cohort, with $a = 1.30, 1.50$, $n = 400$ and 40 runs per value. Full-batch SGD uses learning rate 0.3 and every-step crossing detection. Its 32,000-step budget was fixed before the scored runs by a calibration rule that observed only output-weight growth. The sample thresholds had already been computed. The budget rule aimed for at least 90% crossing; 30 of 40 crossed at each value, exactly the registered minimum. A post hoc sensitivity analysis imputes each non-crosser at its final scale, a lower bound on an unobserved crossing. The own-versus-population test still passes, with intervals $[-0.056, -0.004]$ and $[-0.056, -0.003]$. The correlation test fails at both values: Spearman correlations become 0.502 and 0.507. The SGD own-sample result is therefore a partial pass, holding for crossers only. The

Table 25: Registered SGD predictions, every-step detection. Errors are absolute log errors; intervals are paired 95% intervals. Both own-versus-population and correlation tests pass among crossing runs.

a	Crossed	Error difference (own minus population)	ρ	Own lag	Pop. lag
1.30	30/40	-0.0459 $[-0.0738, -0.0172]$	0.674	2.95%	9.45%
1.50	30/40	-0.0439 $[-0.0727, -0.0148]$	0.684	6.43%	12.74%

non-crossers end at median 0.15 and 0.30 times their own thresholds, respectively, so these are loose lower bounds, not observed crossings. The timescale test below still passes under this imputation, with observed median lags 0.0277 and 0.0586.

L.2 Two prospective tests of a frozen timescale relationship

The post hoc Adam fit is $r = 0.0157 + 2.658\chi$, where r is the lag relative to the own threshold and χ is the ratio of output growth to branch relaxation at crossing. For Adam, the denominator is $\eta_{\text{opt}}\lambda_{\min}(D^{-1/2}HD^{-1/2})$; for SGD the preconditioner is the identity. The fitted relationship was frozen before SGD scoring. Measuring χ requires a trajectory at crossing: the prospective part is the frozen lag relationship, not an advance prediction of that trajectory's ratio. The SGD ratios lie

Table 26: Registered SGD timescale extension, among 30 crossings per activation value, detected every step. Tolerances are acceptance limits, not uncertainty intervals. Both comparisons pass.

a	Median χ	Predicted lag	Observed	Tolerance	Verdict
1.30	0.00354	0.0251	0.0295	± 0.0100	Pass
1.50	0.01404	0.0530	0.0643	± 0.0133	Pass

inside the Adam fit's range. At $a = 1.50$, the miss of 0.0113 is close to the tolerance of 0.0133. This test interpolates across optimizers.

The second test was registered before any run, at $a = 1.45, 1.60$ on fresh samples unused elsewhere. Each sample threshold was frozen by hash before Adam training. The fit had seen only $a = 1.30, 1.50$; every-step detection is used, and all 80 runs per new value cross. The non-crosser

Table 27: Registered fresh-sample test at unseen activation values, 80 crossings per value, every-step detection. The last column gives the paired 95% interval of own minus population prediction error. Every listed test passes. Tolerances are registered acceptance limits.

a	Median χ	Predicted	Observed	Tolerance	Own minus population
1.45	0.0112	0.0456	0.0532	± 0.0114	$[-0.059, -0.022]$
1.60	0.0249	0.0818	0.0809	± 0.0205	$[-0.059, -0.022]$

sensitivity analysis is identical because every run crosses. At $a = 1.60$, the median ratio 0.0249 is slightly beyond the fit's maximum 0.0229, so this prediction is a slight extrapolation. A second registered test, at activation values the fit never saw and on fresh training sets, predicted the median lag from the Adam timescale relationship within tolerance at both values, and each run's own threshold again beat the population threshold. These prospective tests support the frozen fitted timescale relationship. The no-fit tracking calculation was derived later, after the fitted relationship was known. It subsequently received a separate prospective test at $a = 1.65$ (Appendix V.7).

M Timescale relationship and within-activation interventions

The initial timescale analysis is post hoc. It replays 384 crossings from the deconfounded growth-rate test, reproducing every crossing exactly. The pooled Spearman correlation of lag with growth divided by relaxation is 0.634 $[0.542, 0.721]$; within $a = 1.30, 1.50$ it is 0.454 and 0.507. Across the eight cells it is 1.0. Output growth varies only about 25% across an eightfold output-learning-rate range because Adam's normalization compensates. The pooled association partly reflects differences between activation values.

A later post hoc check applies the frozen relationship to 1,530 crossing runs from the original growth-rate test, both timing rules of the deconfounded test and the lag interventions. Every replay reproduces its recorded crossing exactly. Predictions are evaluated at the prospective tests' tolerance $\max(0.01, 0.25|\text{prediction}|)$. Table 28 gives all arms. The relationship predicts 28 of 32 within tolerance. The four misses are the strongest slowdowns applied from initialization or plateau exit at both activation values; the predicted direction is right, but the lag falls further than predicted. All arms that leave the ratio near its unperturbed value are within tolerance.

Within-activation slopes are 4.69 at $a = 1.30$ and 3.09 at $a = 1.50$, versus 2.66 pooled. Spearman partial correlation controlling for activation is 0.51 across all arms and 0.48 on the fit's data. Thus the association is not solely an effect of comparing activation values, but the pooled linear fit understates the effect of large within-activation changes.

A single linear relation between the lag and the ratio of output growth to branch relaxation predicts the lag across activation values and optimizers (two registered prospective tests) and for within-activation interventions that leave the ratio nearly unchanged. It captures the direction but underestimates the size of the effect when the output rate is slowed substantially (post hoc). The registered proportionality and final-approach predictions remain failed. This relationship is not an explanation of width-2 training.

Table 28: Post hoc application of the frozen timescale relation to all 32 intervention arms, with every-step crossings. The tolerance is an acceptance limit, not an uncertainty interval. The lag-intervention rows reproduce the original arms as run; their corrected reruns are disclosed in Appendix H. "prim" starts at 0.7 of the own threshold; "tstar" starts at plateau exit.

Test	a	Arm	n	Ratio	Predicted	Observed	Tolerance	Within
4b	1.30	control	48	0.0040	0.0263	0.0311	±0.0100	yes
4b	1.30	reset	48	0.0041	0.0266	0.0315	±0.0100	yes
4b	1.30	teleport	48	0.0040	0.0264	0.0308	±0.0100	yes
4b	1.30	teleport_reset	48	0.0041	0.0266	0.0312	±0.0100	yes
4b	1.50	control	48	0.0157	0.0574	0.0656	±0.0143	yes
4b	1.50	reset	48	0.0163	0.0591	0.0672	±0.0148	yes
4b	1.50	teleport	48	0.0158	0.0576	0.0655	±0.0144	yes
4b	1.50	teleport_reset	48	0.0161	0.0585	0.0660	±0.0146	yes
lag1	1.30	0.25	49	0.0009	0.0181	0.0047	±0.0100	no
lag1	1.30	0.5	48	0.0035	0.0250	0.0266	±0.0100	yes
lag1	1.30	1.0	48	0.0040	0.0263	0.0311	±0.0100	yes
lag1	1.30	2.0	44	0.0040	0.0264	0.0312	±0.0100	yes
lag1	1.50	0.25	49	0.0035	0.0249	0.0107	±0.0100	no
lag1	1.50	0.5	48	0.0146	0.0545	0.0534	±0.0136	yes
lag1	1.50	1.0	48	0.0157	0.0574	0.0656	±0.0143	yes
lag1	1.50	2.0	44	0.0158	0.0576	0.0664	±0.0144	yes
lag2-prim	1.30	0.25	48	0.0034	0.0248	0.0278	±0.0100	yes
lag2-prim	1.30	0.5	48	0.0038	0.0259	0.0302	±0.0100	yes
lag2-prim	1.30	1.0	48	0.0040	0.0263	0.0311	±0.0100	yes
lag2-prim	1.30	2.0	48	0.0041	0.0265	0.0316	±0.0100	yes
lag2-prim	1.50	0.25	48	0.0135	0.0515	0.0563	±0.0129	yes
lag2-prim	1.50	0.5	48	0.0150	0.0556	0.0629	±0.0139	yes
lag2-prim	1.50	1.0	48	0.0157	0.0574	0.0656	±0.0143	yes
lag2-prim	1.50	2.0	48	0.0162	0.0587	0.0667	±0.0147	yes
lag2-tstar	1.30	0.25	48	0.0013	0.0192	0.0071	±0.0100	no
lag2-tstar	1.30	0.5	48	0.0035	0.0251	0.0271	±0.0100	yes
lag2-tstar	1.30	1.0	48	0.0040	0.0263	0.0311	±0.0100	yes
lag2-tstar	1.30	2.0	48	0.0041	0.0265	0.0315	±0.0100	yes
lag2-tstar	1.50	0.25	48	0.0054	0.0300	0.0164	±0.0100	no
lag2-tstar	1.50	0.5	48	0.0146	0.0544	0.0557	±0.0136	yes
lag2-tstar	1.50	1.0	48	0.0157	0.0574	0.0656	±0.0143	yes
lag2-tstar	1.50	2.0	48	0.0162	0.0587	0.0674	±0.0147	yes

N Additional width-2 tests

N.1 Asymmetric windows: a conditional threshold and a failed training prediction

The registered asymmetric task uses $I = [-0.8, 0.8]$, $O = [-2.0, -1.2] \cup [1.2, 2.4]$, and $a = 1.30$. Here $\mathbb{E}_O x - \mathbb{E}_I x = 0.44$, so the ramp contributes to the class-mean gap. The small-scale criterion predicts a conditional placement switch. A validated search finds unplaced minima at $s \leq 0.4217$ and placed minima at $s \geq 0.5623$ on a 17-scale grid, with bisection bracket $s \in [0.4371, 0.4532]$. Restart counts from 500 to 4,000 leave it unchanged, independent CMA-ES agrees and the audit is

clean. The registered threshold and secondary comparison pass. These are validated searches, not global certificates.

The registered width-2 training prediction on asymmetric windows, T2-3, failed: runs crossed at a median 3.3 times the threshold. Two follow-ups were registered after that failure, each labelled as such, to put width-2 training in width 1's timescale regime. T2-3b (steps-matched output rate) is unresolved: its runs crossed at a timescale ratio far below width 1's range. T2-3c (ratio-matched at the threshold scale) is unresolved for the same reason: the ratio at crossing fell far below the range. Neither follow-up tested the prediction, and T2-3 stands as a failure. A standard-initialization arm

Table 29: Registered asymmetric width-2 training tests, in registration order. Ratios compare crossings with the lower conditional-threshold bracket endpoint. Follow-up validity failures prevent a verdict on their outcome predictions.

Test	Design	Result	Verdict
T2-3	Matched initialization; 80 runs	One initially placed run excluded; 79 crossed. 93.7% cross at or above the upper endpoint; median ratio 3.29 exceeds 1.25. Interval for median ratio minus one: $[1.89, 2.56]$.	Fail
T2-3b	Output-rate factor 0.0093, chosen by steps matching	Crossing timescale ratio 6.8×10^{-6} lies below the required $[0.0014, 0.023]$.	Unresolved
T2-3c	Output-rate factor 0.01778, chosen by threshold-scale ratio matching	Crossing timescale ratio 1.3×10^{-5} lies below the same range.	Unresolved

ran before the matched-initialization amendment; its outputs were sealed unread with committed hashes and are withdrawn and unscored. For information only, T2-3b had 71 of 80 crossings, 56.3% above the upper endpoint, median ratio 10.4 and interval $[-0.41, 9.52]$; T2-3c had 80 of 80 crossings, 47 of 80 above the endpoint and median ratio 10.8. These numbers do not override the failed validity checks. The latter pilot observed only the ratio when scale first reached the threshold, never placement; the registered risk that crossings occur much later materialized.

Post hoc diagnosis. The registered training prediction failed. Post hoc, the runs cross as cooperating pairs, on the branch whose switch is the width-2 threshold, but about three times above it, and below the single-unit (width-1) threshold. Neither threshold, nor the branch's own switch, predicts the crossing scale. Of the 79 original crossings, 76 are cooperating pairs and three are redundant; none is single-unit. Their median crossing scale is 1.44, compared with the width-1 threshold 5.09 and median occupied-branch switch 0.439. The branch-minus-global prediction-error interval $[-0.196, +0.015]$ includes zero.

An exploratory own-sample check on ten seeds gives median width-2 threshold 0.445, range 0.346–0.573. Runs cross at 4.1 times their own threshold, versus 4.3 times the population threshold; Spearman correlation is -0.74. The median crossing timescale ratio is 2.73, far outside the width-1 range. Extrapolating the width-1 fit predicts lag 7.3 versus 2.23 observed. These checks do not account for the width-2 lag.

A further post hoc diagnostic gate compared the two slowed arms with the original arm. The best pooled predictor was continuation from the branch occupied at crossing: median absolute log errors 0.010 and 0.006 in the slowed arms, but 1.17 in the original arm, so the gate stopped. This branch predictor uses the crossing configuration and is close to circular; it is undefined for 41 slowed-arm runs. Own-sample width-1 thresholds gave errors 0.015 and 0.011 in ten seeds per slowed arm. This observation did not explain the original training result.

Prospective test restricted to slowed growth. A registered test designed after the diagnosis found that, in the slowed regime, the median run crosses at its own width-1 threshold, but about 40% of runs cross early at small scale, so the registered comparison with the width-2 threshold fails. The full-speed regime lies outside the measured tracking range; the post hoc branch analysis is reported in Appendix X. This test uses output-rate factor 0.01778 and fresh samples, with each width-1 own-sample threshold frozen by hash before training. Median absolute log error is 0.019 and the median

crossing is 0.98 times that threshold, satisfying the accuracy criterion. The paired interval against the population width-2 predictor is $[-1.04, +0.0064]$, so the comparison fails. Of 79 crossings, 47 occur at the width-1 own threshold and 31 occur early at $s = 0.016$–0.42. The registered failure does not establish a width-1-threshold predictor for width-2 training.

N.2 Symmetric windows: fixed-scale training still shows gating

The registered no-gating prediction tests width 2 at $a = 1.30, 1.50$. Latest pre-placement checkpoints of 80 runs per value are replayed at held $R_2 \in \{0.003, 0.01, 0.03, 0.1\}$. A pilot unable to see the outcome fixed the horizon at 16,000 steps. The primary arm preserves the optimizer state; placement uses continuous-window extrema. The positive control passes: width 1 gives no placement at $R/R_{\text{glob}} = 0.1$ at either activation value. At the horizon, unplaced endpoints are stationary two-unit configurations, with median scale-relative gradient 1.4×10^{-7} and 59% below the 10^{-6} tolerance. None is a single-unit local minimum, so the registered longer-horizon extension applies to no runs. Only five of 640 endpoints change status between one-quarter of the horizon and its end, all to placed; placed endpoints are almost all cancelling pairs.

Table 30: Placed fractions in the symmetric width-2 test at the frozen horizon. Each f_a row has 80 checkpoints per held scale. These fractions are descriptive point estimates; no uncertainty intervals are supplied. The registered no-gating prediction fails at both activation values.

Arm	$R_2 = .003$	.01	.03	.1
$a = 1.30$, preserved	0.46	0.50	0.68	0.99
$a = 1.50$, preserved	0.46	0.49	0.64	0.97
$a = 1.30$, reset	0.44	0.45	0.64	1.00
$a = 1.50$, reset	0.42	0.46	0.62	0.97
tanh, preserved (descriptive)	0.59	0.61	0.68	0.75

The landscape predicts no gating at width 2 on symmetric windows, and none exists in the conditional minimizer at the searched scales. The registered training test nevertheless found gating: at small held scale about half the runs remain unplaced, at stationary two-unit configurations. So the conditional-threshold account does not by itself predict width-2 training.

O Budget dependence

Budget dependence under Adam. If $\widehat{G} \propto \varepsilon^{\beta}$ and attained output scale grows as B^{α} without further ε-dependence near the onset, a fixed transition value of R gives $\varepsilon_{\text{on}} \propto B^{-\alpha/\beta}$. Over budgets 10^3–1.6×10^5 at $a = 1.25$, the measured effective weight-growth exponent is $\alpha = 1.1172 \pm 0.0605$. The registered onset prediction was -0.7448; the measured exponent is -0.7340 ± 0.0370, with all six onsets bracketed.

The separability assumption is approximate. A fit allowing $|w_2| \propto B^{\alpha}\varepsilon^{-\nu}$ gives $\nu = -0.124$ with interval $[-0.238, -0.029]$, failing the registered zero-dependence test. The materiality check passes: its exponent correction is 0.057 against grid resolution 0.114. Weight-growth exponents also vary systematically with ε (correlation -0.9943), so the separable fit is misspecified. Across budget windows, α falls from 1.5109 on 10^3–4×10^3 to 0.7193 on 4×10^4–1.6×10^5. Held-out extrapolation fails: a fit to the four smallest budgets predicts $\varepsilon = 0.01872$ with interval $[0.01403, 0.02497]$, while the measured value is 0.030. We use the law for interpolation over the measured budgets.

P Optimizer equivalence and pooled prediction

The equivalence experiment targets the transition band at $a = 1.25$. A placement pilot selected budget cells, after which the registered design collected at least 500 in-band runs per optimizer. The band-targeted $a = 1.25$ midpoint estimates are maximum-likelihood logistic fits. Intervals for pairwise differences use 4,000 cell-level bootstrap resamples; two-one-sided tests use equivalence margin $\delta = 0.024$.

Table 31: Band-targeted $a = 1.25$ population only.

Optimizer	Runs	In band	Band-targeted R_{50}
Adam	1,200	708	0.3471
AdamW	960	687	0.3479
SGD	840	653	0.3442

Table 32: Pairwise midpoint differences in the band-targeted $a = 1.25$ population. Each interval lies inside the equivalence margin.

Pair	Difference	90% interval
Adam − AdamW	-0.0008	$[-0.0052, +0.0043]$
Adam − SGD	$+0.0029$	$[-0.0040, +0.0086]$
AdamW − SGD	$+0.0037$	$[-0.0038, +0.0092]$

The registered expectation that comparisons involving SGD would be non-equivalent or inconclusive failed. This equivalence concerns the band-targeted midpoint, not the budget law or the branch mechanism. Under SGD the proposed onset relationship fails.

The pooled midpoint $R_{50} = 0.3682\,[0.3603, 0.3730]$ comes from a binned logistic fit to 3,150 untargeted runs across activation parameters and budgets. It has a different sampling design from the band-targeted experiment. We do not interpret the difference between these population-specific estimates.

At fixed a, $\widehat{G}(a)$ is constant, so R ranks runs identically to $|w_2|$. The product's contribution is alignment across activation parameters. The exponent scan over $|w_2|^p \widehat{G}^q$ selected ratios p/q between 1.00 and 1.18 in the three families where the relationship holds, with maximum improvement 0.0008 AUC. Lag activation-parameter effects remain: in two nearby R windows the reported coefficients have $z = -2.53$ and $z = -1.92$. These measurements do not support sufficiency of R.

The product aligns outcomes across activation parameters much better than either factor alone. In the pooled population of 3,150 untargeted runs, no run below $R = 0.30$ solves (0 of 2,281), while 461 of 463 above $R = 0.50$ solve. Both exceptions occur under Adam at $a = 3.0$ after 2,000 steps. One has positive gap but misses the allowable bias interval by 0.010958. The other has $w_1 = 0.019358$ and negative gap, despite $R = 0.500886$. A large geometry factor can raise R even when the hidden unit is nearly constant. Equation (2) distinguishes these failures.

Q Activation controls and budget sensitivity

For $q > 0$, the constructed activation is

$$f_{a,q}(t) = (1 - a)t + \frac{a}{q+1}\,\mathrm{sign}(t)|t|^{q+1}, \qquad |t| \le 1,$$

continued continuously with slope one outside this interval. Its local dip exponent is $\beta = 1 + 1/q$, verified numerically within 0.7%. The scaling reduction for a polynomial segment applies only when all relevant preactivations remain inside that segment. The cross-family follow-up used a scale-equivariant conditional-search procedure after the frozen procedure failed its small-ε validity gate. This procedural deviation must accompany the successful follow-up.

The Adam-only geometric onset comparison gives a through-origin slope of 1.0984 with interval $[0.958, 1.239]$. Replacing asymptotic exponents with effective exponents gives 1.0547. The effective family-A dip exponent is 1.4269 over $\varepsilon \in [0.025, 0.6]$. The registered collapse-selectivity criterion fails, and changing the solve-rate threshold affects bracketing coverage. Neither check strengthens the held-out extrapolation, which fails.

Homogeneous family B. Let $f_\alpha(t) = \max(t, \alpha t)$ with $\alpha < 0$. Positive homogeneity gives $G(cw_1, cb_1) = cG(w_1, b_1)$ for $c > 0$, so the unrestricted supremum is infinite whenever a positive

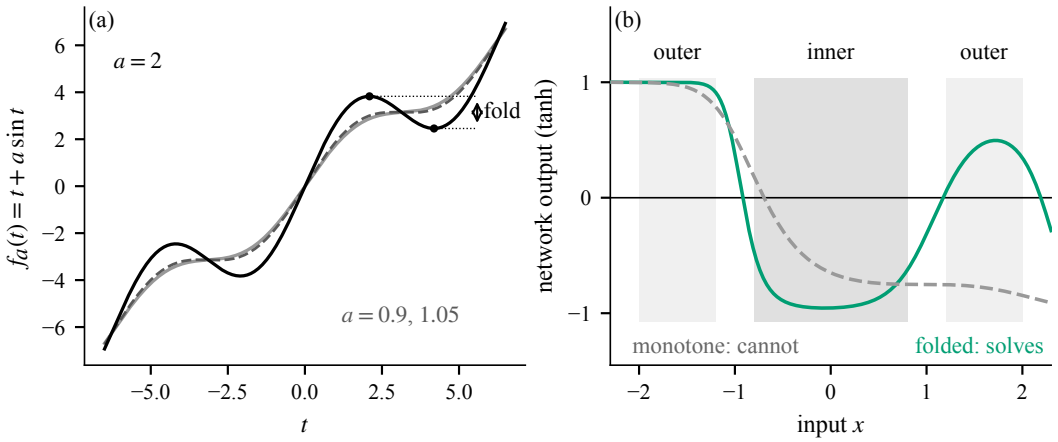


Figure 13: For $a > 1$, a non-monotone unit can put both outer windows above the inner window, whereas a monotone unit cannot. Panel (a) shows the closed-form activation at $a = 0.9, 1.05, 2.0$, with $D(2) = 1.3697$. Panel (b) shows two individual float64 width-1 runs trained with Adam at learning rate 0.01 for 2,000 steps: $a = 1.5$, seed 0, solves a dense 4,001-point check; $a = 1.0$, seed 38, is the best monotone run, with 91 of 200 sample errors. Outputs are shown through tanh. The windows are $I = [-0.8, 0.8]$ and $O = \pm[1.2, 2.0]$. No uncertainty is shown. This setting illustration complements the main-text mechanism figure.

gap exists. Family B is outside the finite-$G^\star$ construction. Its normalized comparison fixes $|w_1| = 1$ and uses $\widehat{G}_{\mathrm{norm}}(\alpha) = 0.4|\alpha|$ for the tested values $\alpha \in \{-1.0, -0.5, -0.25, -0.1, -0.05\}$. The normalized supremum ranges over b_1 and the input-weight orientations with $|w_1| = 1$. The corresponding scale variable uses $|w_1 w_2|\widehat{G}_{\mathrm{norm}}/2$.

Under this normalization the geometric exponent is exactly $\beta_B = 1$. The corrected required-scale overshoot is 11.8–145 times. The registered range prediction omitted a budget row and failed. The formal budget-onset prediction is -1.1172, while the reported measured exponent is $+0.25$; the positive sign is unaffected by the normalization. The AUC comparison has been removed because its design cannot establish the proposed claim. Within a fixed α cell the normalized product is a constant multiple of $|w_1 w_2|$, and the transition-band coverage was inadequate in most cells.

For $q = 2$, the limiting polynomial is $-z + z^3/3$. Substitution gives $h(\sqrt{2}z) = \sqrt{2}(-z + z^3/3)$, deriving the observed $\sqrt{2}$ prefactor relation. The first registered cross-family test failed: at $a = 1.30$–1.60, placement thresholds agreed but solve thresholds did not, with 25–34% of window points outside the constructed family's polynomial domain. A separately registered follow-up at $a = 1.01$–1.04 kept every window point inside that domain. Both thresholds agreed at all tested values to grid resolution 0.00215. The subsequently registered $q = 1$ control, which has a different limit, differed by 35–36 grid steps on the placement threshold and by 5 steps on the solve threshold. The follow-up supports the reduction within its domain; the first prediction remains failed.

Task geometry. Across the base task and four new window geometries, the cubic prediction matches the measured geometric constant within 0.808%, over an 8.1-fold range of ν_0. No constant is fitted per window. The placement threshold itself is geometry-specific. Comparing each geometry's measured crossing with its then-available conditional estimate meets the registered 15% tolerance in only 6 of 10 cells, so that prediction fails. Dividing each ratio by the base task's ratio at the same a puts all cells within 2.72% of unity. This normalization was chosen after the failures and is reported only as a post hoc observation.

R Measurements at constructed solutions

At $a = 1.02$, 200 Adam runs at 2,000 steps do not solve the width-1 task, although a constructed unit-output-weight classifier does. This is a sweep result, not a dedicated search over optimiza-

tion methods. The following measurements reject specific proposed explanations at the measured configurations.

Table 33: Representative constructed and Adam-found configurations. Distance is Euclidean distance from initialization. The path column is omitted because the discretized path check is described separately.

Population	a	$\lvert w_2\rvert$	$\lambda_{\max}\eta$	Distance	Margin
Constructed	1.02	1.000	0.02718	4.671	0.000274
Constructed	1.25	1.000	0.02775	4.943	0.012929
Constructed	1.50	1.000	0.02914	5.074	0.038420
Constructed	3.00	1.000	0.02477	5.517	0.332924
Adam-found	1.50	4.394	0.002800	14.199	0.077061
Adam-found	3.00	3.472	0.000517	10.232	1.643642

The recorded found configurations are the first solving seeds among seeds 0–39 under full-batch Adam at learning rate 0.01 and 2,000 steps.

The path calculation found a discretized path whose evaluated images never exceed the higher endpoint loss. It was not a converged minimum-energy path, and finite image values do not certify the loss between images. Each path string contains 41 points. The calculation uses 400 update steps at learning rate 0.005, fixed endpoints and five initializations for each of 12 endpoint pairs. The zero sampled barrier holds across those 60 strings. The constructed configurations are below the full-batch gradient-descent reference $\lambda_{\max}\eta = 2$ by at least a factor of 66. This reference is not an Adam stability theorem. The registered sharpness explanation fails for these configurations.

The constructed points are nonstationary: gradient norms are 0.0192–0.2724, and the minimum Hessian eigenvalue is negative at six of eight points. We initially checked only the sharpness condition in Ahn et al. (2022); nonstationarity is the more basic reason its stationary-point result does not apply.

The pathwise check compares 400 trajectories with verified constructed solutions, measuring Euclidean distance from the true initialization. In 114 of 400 runs, the trajectory never gets closer than its starting point. The largest approach is 23.4% closer. Runs end about ten times farther away than they start, with median distances 49.7 and 5.2. At $a = 1.5$, 17 of 81 float32 Adam-found solutions have margins below the construction's 0.0384, with minimum 0.0017. Margin therefore does not order these populations. The representative found row in the table has a larger margin.

S Continuous errors and metric choice

Metric choice. The metric concern of Schaeffer et al. (2023) motivates the continuous-error check. In its 2,400-run population, 76% of the error improvement occurs while no run solves. Under the certified geometry, the steepest continuous-error slope is $-1{,}760$, compared with -142 at the binary threshold. The crossing bands depend on binning and overlap at the figure's bin width. The registered width-ratio verdict was intermediate. These results prompted withdrawal of a sharp capability-threshold interpretation. Achieved margin is not independent evidence, since its sign defines success.

T Further related work

Topology constrains narrow networks' class arrangements (Naitzat et al., 2020; Guss & Salakhutdinov, 2018; Ren & Lim, 2026); minimum-width results identify approximation obstructions (Johnson, 2019; Hanin & Sellke, 2017; Park et al., 2021; Cai, 2023; Kim et al., 2024; Kidger & Lyons, 2020; Li et al., 2023). Continuous-depth flows have related restrictions that state augmentation can remove (Chen et al., 2018; Dupont et al., 2019). Depth separation and linear-region complexity measure representation rather than trained outcomes (Telgarsky, 2016; Eldan & Shamir, 2016; Montúfar et al., 2014; Raghu et al., 2017); expressible targets that resist gradient learning are already known (Shalev-Shwartz et al., 2017). Non-monotone and periodic activations have other motivations

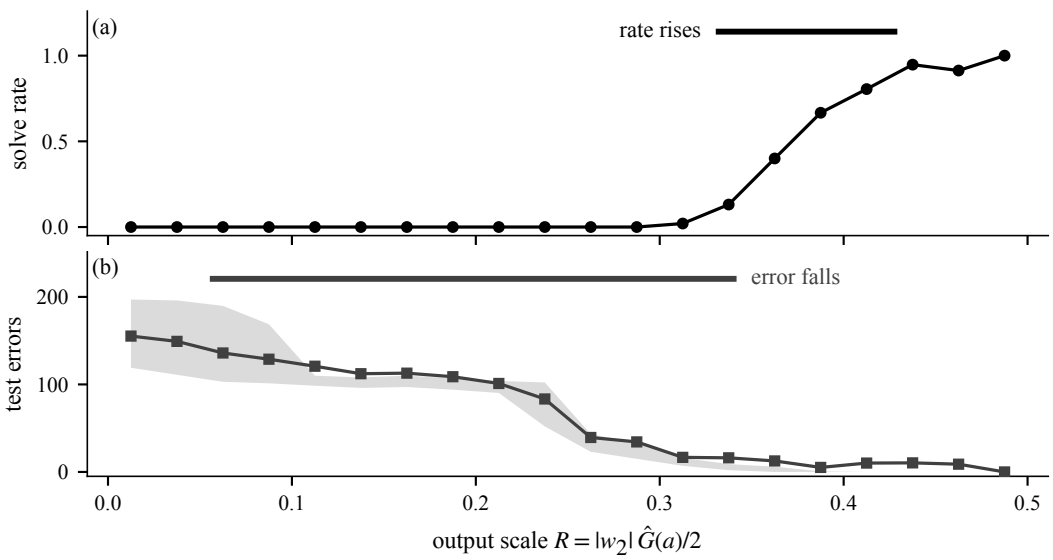


Figure 14: Continuous test error falls while binary solve rate remains zero. This metric detail is omitted from the main-text placement figures. There are 2,400 width-1 runs, binned by R at width 0.025. Panel (a) gives solve fractions; panel (b) gives mean errors with an interquartile band, the only uncertainty shown. The 10–90% binary transition spans $R \in [0.330, 0.429]$; the corresponding continuous-error fall spans $[0.055, 0.342]$, a width ratio of 2.9. Of the error improvement, 76% occurs in zero-solve bins (1,815 runs, $R < 0.30$). The binary transition does not establish a sharp capability onset.

(Hendrycks & Gimpel, 2016; Ramachandran et al., 2017; Misra, 2020; Sitzmann et al., 2020; Ziyin et al., 2020); our tunable dip permits an analytic boundary and scaling reduction.

Metric choice can manufacture apparent emergence (Schaeffer et al., 2023): here 76% of error improvement occurs while no run solves, so the binary solve transition is not a sharp capability onset. Compute-scaling work (Kaplan et al., 2020) concerns a different quantity from our budget-dependent placement onset, whose effective exponent does not extrapolate.

Minimal models permit analysis of edge-of-stability dynamics (Zhu et al., 2023); the constructed-solution measurements distinguish that proposed explanation from the observed delay (Cohen et al., 2021; Ahn et al., 2022). A threshold-neuron model derives a transition from the dynamics (Ahn et al., 2023). Teacher–student phase transitions (Rubin et al., 2024), train–test norm mismatch (Liu et al., 2023) and numerical collapse from logit growth (Prieto et al., 2025) concern different mechanisms. Progress measures reveal change before generalization in modular addition (Nanda et al., 2023), and hidden progress also occurs in parity learning (Barak et al., 2022). Margin growth on separable data (Soudry et al., 2018) and homogeneous networks (Lyu & Li, 2020) provides the large-scale context for the present non-homogeneous unit.

U CENTRAL FAILED PREDICTIONS

The relevance classification is post hoc and fixed by topic, never outcome. There are 33 central failed predictions: 31 under registered rules and 2 by post hoc scoring. Table 34 groups them; the supplementary lists the central failures individually and gives the full by-block census. Prediction-level rules are in the accompanying experimental artifacts. The full census has 236 predictions: 214 scored by registered rules (100 passed, 66 failed, 8 partial, 40 unresolved) and 22 assigned post hoc (3 passed, 6 failed, 13 partial). A separate width-2 decision rule is not counted as a prediction.

Table 34: Central failures, grouped without changing their verdicts. The relevance classification is post hoc.

Topic	Claims limited	Failed
Threshold and lag	Placement quality; crossing references; held-scale outcomes; horizon, sample-size and intervention predictions; ramps; practical activations; higher-dimensional lag range; Adam individual ordering.	23
Scaling	Cross-family solve threshold; geometric magnitude/containment; arithmetic and margin subclaims.	4
Prospective geometry	Earlier window prediction and far-outer-window subclaim.	2
Width two	Tanh boundary criterion; full-speed training; no-gating; slowed own-threshold comparison.	4

V THE LAG LAW AND ITS LINEAR-RESPONSE REFINEMENT

V.1 DERIVATION AND ASSUMPTIONS

This calculation was derived after the fitted timescale relationship was known. It is post hoc and changes no registered verdict. The constant and every prediction were committed before comparison with observed crossings. It assumes one tracked stationary branch, a positive-definite hidden-coordinate Hessian, and a displacement small enough to linearize the gradient and gap.

The hidden coordinates are $\theta = (w_1, b_1, b_2)$. In particular, the output bias is trained, not profiled during the dynamics. At fixed $s = |w_2|$, let $\theta^*(s)$ satisfy $\nabla_\theta L = 0$. Its tangent is

$$\theta^{*\prime} = -H^{-1}\partial_s \nabla_\theta L.$$

Writing $\theta = \theta^* + \delta$ and linearizing a preconditioned gradient step gives

$$\delta_{t+1} - \delta_t = -\eta P H \delta_t - \theta^{*\prime}\dot{s}.$$

If the drift varies slowly on the relaxation timescale, the steady displacement is $\delta = -(\eta P H)^{-1}\theta^{*\prime}\dot{s}$. Since the continuous-window gap is independent of b_2, linearizing its zero at s^* gives

$$(\nabla G^\top \theta^{*\prime})(s_c - s^*) = -\nabla G^\top \delta,$$

which yields Equation (7). Scaling P by a constant leaves κ invariant. Its shape matters, while its size enters χ. The original comparison measures $\dot{s}/s_c$ at crossing, rather than $\dot{s}/s^*$; the difference is the factor $1 + r$ and is second order in the lag.

For Adam, $P = \mathrm{diag}(1/(\sqrt{\widehat{v}} + \epsilon_{\mathrm{opt}}))$. The calculation treats it as fixed over relaxation and uses the median normalized preconditioner in the existing crossing population. Steady displacement makes the gradient constant, so Adam's first moment equals that gradient. Momentum changes transients but not the steady lag. Its time constant is 10 steps; the measured relaxation times are listed below. A nearly singular Hessian, fast scale growth, a rapidly changing preconditioner or the wrong branch invalidates the approximation.

Winding and mirror symmetries. The transformation $(w_1, b_1, b_2) \mapsto (w_1, b_1 + 2\pi k, b_2 - 2\pi k s)$ preserves logits at fixed scale. It changes the tangent by $-2\pi k e_{b_2}$, and therefore changes κ. Profiling the bias during training would remove this effect, but is not the protocol. All 1,750 predicted runs use winding $k = -1$ at $a = 1.30$ and $k = 0$ at the other values. Reflection preserves the population constant; 875 runs occupy the mirror branch. The gap gradient uses centered differences with step 10^{-6}. One-sided estimates agree within 5×10^{-6} relative. The interquartile range of κ over individual preconditioners lies within 0.9% of the median. These are numerical calculations, not interval certificates.

V.2 CLOSED-FORM COMPARISON

All 36 arms meet the tolerance $\max(0.01, 0.25|\mathrm{prediction}|)$, with no refit. The original lag reference is the global own-sample threshold. Through-origin observed/predicted slopes are 0.89, 0.83, 0.91

Table 35: No-fit lag constants computed on the 800-point population branch. Every switch lies within its certified placement bracket. Adam relaxation times use the absolute crossing preconditioner; the constant uses its relative shape.

a	s^*	k	κ_{Adam}	κ_{SGD}	Relaxation steps
1.30	4.9580	-1	7.61	7.62	7.8
1.45	2.8962	0	4.59	4.61	12.1
1.50	2.5270	0	4.05	4.07	14.7
1.60	2.0022	0	3.29	3.32	16.2

and 0.82 in increasing activation order. Four strong-slowdown arms have ratios 0.70–0.76 and meet tolerance only through its absolute floor. Thus agreement of every arm does not mean uniform relative accuracy. Against the tracked-branch switch, arm-median ratios are 0.91–1.09.

V.3 Trajectory-integrated linear response

The refinement is post hoc, developed after the closed form's shortfall was known. Predictions were committed and hashed before comparison. Every run was replayed deterministically under its original protocol, recording output scale and optimizer state without evaluating the actual trajectory's placement gap. Conditional branches use each run's own sample. Newton continuation defines the branch and its exact-extrema switch; Hermite interpolation supplies the branch path and linear interpolation the Hessian.

With Adam's first moment retained as a state, the recursion is

$$m_{t+1} = \beta_1 m_t + (1-\beta_1) H(s_t)\delta_t,$$
$$\delta_{t+1} = \delta_t - \eta P_{t+1} \frac{m_{t+1}}{1-\beta_1^{k_{t+1}}} - \big[\theta^*(s_{t+1}) - \theta^*(s_t)\big].$$

For SGD it reduces to $(I - \eta H)\delta_t - \Delta\theta^*$. Initialization uses the observed displacement and first moment at the specified start. An intervention inside the interval is applied as the observed parameter and moment jump. The prediction is the first step with $G(\theta^*(s_t) + \delta_t) > 0$. The committed stability rule gives no prediction when the iterated one-step map has spectral radius above unity or the displacement exceeds its allowed bound.

All 1,750 runs were processed. No run was on a different branch at either start scale, $0.7s^*$ or $0.5s^*$. None was excluded for branch mismatch. The stability rule leaves 42 predictions undefined from the first start and 23 from the second. Both predictions exist for 1,706 runs and give identical crossing steps in every case. Most of the first-start exclusions are reset arms whose temporarily tiny second-moment estimate makes the linearized update unstable. The full model matches closely.

Table 36: Post hoc linear-response comparison. Values are through-origin slopes of observed on predicted lag; parentheses give median per-run ratios. The first column of predictions uses the global own-sample reference, the others the tracked-branch switch.

a	Runs	Closed form, global	Full, from $0.7s^*$	Full, from $0.5s^*$
1.30	795	0.89 (1.02)	1.02 (1.02)	1.02 (1.02)
1.45	80	0.83 (1.01)	1.03 (1.03)	1.03 (1.03)
1.50	795	0.91 (1.00)	1.04 (1.04)	1.04 (1.04)
1.60	80	0.82 (1.01)	1.05 (1.05)	1.05 (1.05)

Observed/predicted ratios of arm medians are 1.00–1.05 in every arm, from either start, over runs with a defined prediction. The lag is quantitatively first-order tracking in this population. The analysis uses measured scale paths and optimizer states, so it is not a forecast of an entire trajectory from initialization.

The committed prediction-side implementation chose stability checks and ablation definitions before comparison. During development, an exact-gradient reconstruction was run on two trajectories and

therefore exposed their crossing scales. No setting was tuned on them, and that reconstruction was removed from the committed prediction variants. This disclosure limits the claim to a post hoc analysis with predictions committed before its comparison.

V.4 What accounts for the closed form's shortfall

Table 37: Post hoc ablations: observed-on-predicted through-origin slopes. Ablation rows omit the 96 reset runs only; their full-model results remain in the overall comparison.

Reference and model	1.30	1.45	1.50	1.60
Global own threshold, closed form	0.89	0.83	0.91	0.82
Tracked-branch switch, closed form	1.02	1.01	0.98	0.90
Own-switch coefficients, steady displacement	1.01	0.99	0.96	0.90
Frozen coefficients, actual scale path	1.01	1.02	1.03	1.04
Time-varying preconditioner only	1.02	1.03	1.04	1.04
Time-varying Hessian and branch, fixed preconditioner	1.01	1.02	1.03	1.04
All time-varying, no momentum, where stable	1.02	1.03	1.03	1.03
Full model	1.02	1.03	1.04	1.05

The reference explains most of the difference at the smaller activation values: 22.8% of runs track a branch whose switch differs by more than 1% from the global own threshold. At the larger values, the steady-state approximation overpredicts the largest lags. Iterating even the frozen-coefficient model along the actual scale path removes that shortfall. Interpreting this as variation of output growth within the relaxation memory is consistent with the ablation, rather than a separately tested mechanism.

Time-varying Hessians, tangents and preconditioners move the slope by at most 0.01. Momentum leaves the predicted lag unchanged where the no-momentum recursion is stable, but without momentum most linearized Adam paths have no stable prediction. The full model still underpredicts by 2–5%. Replacing its linearized gradient by the exact gradient reproduces every observed crossing step in the checked run-starts, locating that remaining error in nonlinear response. The exact-gradient check covers 856 run-starts and was performed after the comparison.

V.5 Input provenance and relative accuracy

The constant uses no fitted coefficient. It nevertheless has measured inputs, as Table 38 makes explicit. The later registered test freezes measurement rules before training; it does not turn run-dependent preconditioners or output-scale paths into landscape-only inputs.

Table 38: Inputs to the original lag calculation. The final row enters the predictor through χ, not κ.

Input	Origin	Computation
H	Landscape	Hessian in (w_1, b_1, b_2) at the switch.
$\theta^{*\prime}$	Landscape	$-H^{-1}\partial_s \nabla L$.
∇G	Landscape	Central differences of the exact-extrema gap on the active pair.
s^*	Landscape	Gap root on the certified branch, inside the certified bracket.
Winding	Measured runs	Hidden bias at crossing relative to the branch representative.
Adam P shape	Measured runs	Coordinatewise median crossing preconditioner over 48–80 existing runs per activation value.
SGD P	Fixed	Identity.
χ	Measured runs	Output growth and each run's metric at crossing.

The lag-figure intervals use run-level bootstrap resampling within each arm: 10,000 resamples, seed 20260926, 95% intervals of medians. They are sampling intervals, distinct from deterministic threshold enclosures.

Table 39: Relative accuracy depends on the lag reference. Percentages describe individual runs; counts describe arm medians.

Model and reference	Runs	$\pm 10\%$	$\pm 20\%$	Arms $\pm 10\%$	Arms $\pm 20\%$
Closed form, tracked branch	1,750	88.2%	97.8%	36/36	36/36
Closed form, global own threshold	1,750	63.7%	75.4%	31/36	32/36
Integrated, tracked branch	1,708	99.9%	100.0%	36/36	36/36

V.6 Static sample shift and dynamic lag

For each run the identity

$$\log \frac{R_c}{R_{\text{glob}}} = \log \frac{R_{\text{own}}}{R_{\text{glob}}} + \log \frac{R_c}{R_{\text{own}}}$$

separates the static finite-sample shift from the lag behind that sample's threshold. The medians of the terms need not add. With more training data the finite-sample threshold shift shrinks, while the lag behind the own threshold is unchanged.

Table 40: Sample-size decomposition. Brackets are bootstrap 95% intervals of medians.

a	Points	Runs	Total offset	Static shift	Dynamic lag
1.30	400	49	.0682 [.0362,.0935]	.0438 [.0181,.0723]	.0297 [.0273,.0313]
1.30	1,600	48	.0544 [.0311,.0707]	.0396 [.0156,.0522]	.0293 [.0278,.0313]
1.30	6,400	48	.0413 [.0316,.0509]	.0125 [.0057,.0231]	.0306 [.0294,.0312]
1.50	400	49	.1009 [.0800,.1288]	.0447 [.0183,.0726]	.0620 [.0571,.0654]
1.50	1,600	48	.0819 [.0670,.1047]	.0364 [.0135,.0506]	.0619 [.0578,.0658]
1.50	6,400	48	.0721 [.0622,.0801]	.0110 [.0037,.0208]	.0635 [.0602,.0660]

V.7 Registered lag test with no prior crossing data

At $a = 1.65$, there were no existing crossing data. The value had appeared in a landscape table and a deeper-network search, so it was not wholly unused. Eighty fresh seeds per optimizer were registered with the landscape, winding and branch rules, preconditioner rule, per-seed thresholds and prediction pipeline frozen and hashed before training. The population switch $s^*_{\text{pop}} = 1.81002$ is validated by continuation and conditional search, not certified. The retained global minimizer is unplaced at $0.995 s^*_{\text{pop}}$ and placed at $1.005 s^*_{\text{pop}}$ on the same branch.

The winding rule chooses canonical $b_1 \in (-\pi, \pi]$ and was defined from crossings at other activation values, where all 1,750 existing runs obey it. The rule point is the last upward passage of $0.5 s^*_{\text{frozen}}$ before s^*_{frozen}. This departure from the literal first-passage specification was fixed in the registration before any run, because half the threshold lies inside the initialization range. Branch and preconditioner rules use information through this point. The trajectory-integrated model then uses the measured s_t path with P frozen at the rule point. All 160 runs were trained without evaluating their placement gaps; the predictions and path hashes were committed before every-step observation and scoring. No rule point was at or after an observed crossing.

Table 2 reports the scored criteria. Five of six per-optimizer criteria pass; the exception is Adam's per-run rank correlation. The registration census counts the two median criteria jointly across optimizers and the rank criterion separately, giving four prediction rows: three pass and one fail. This convention is distinct from counting the six optimizer-specific outcomes.

Both validity gates require at least 60 crossers and pass. Adam has 76 crossers, with 74 predictions; SGD has 64, with 63 predictions. Four Adam runs and 16 SGD runs never reach the frozen threshold. The branch rule gives no prediction in two Adam crossers and one SGD crosser, and these runs are not replaced. The winding rule agrees with the occupied branch in all 74 predicted Adam runs and all 63 predicted SGD runs.

The trajectory-integrated median per-run observed/predicted ratios are 1.065 for Adam and 1.056 for SGD; closed-form ratios are 1.045 and 1.024. All meet their respective registered bands. The

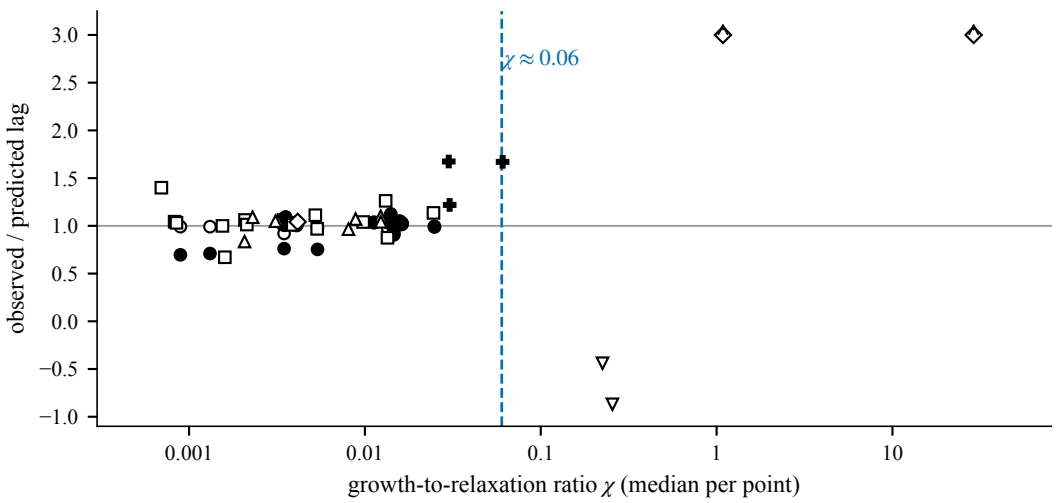


Figure 15: Observed/predicted lag is close to 1 for small ratios and departs as the predicted lag $\kappa\chi$ grows: in the registered SGD boundary test it is 1.22 at $\kappa\chi \approx 0.12$ and 1.67 at $\kappa\chi \approx 0.23$, already at $\chi = 0.03$–0.06. Points are arm or cell medians, with no uncertainty bars: 102 resolved points of 134. Circles denote width-one free training (filled: global own threshold; open: tracked-branch switch), squares forced ramps, upward triangles the band task, downward triangles SiLU and Mish, and diamonds width-two bins. Open markers condition on the branch post hoc. Filled plus signs show registered SGD boundary-test cells with at least 30 of 40 crossings, using the forced-branch switch frozen before any run; cells with fewer crossings are omitted. The earlier branch-conditioned points with $\chi \leq 0.06$ span 0.67–1.40. Width-two ratios reach 33.6 above $\chi \approx 0.3$; SiLU and Mish have the wrong sign at $\chi \approx 0.22$–0.26. Points with $|\text{predicted lag}| < 0.005$ are omitted; ratios above 3 are clipped with arrows.

trajectory-integrated rank correlations are 0.25 and 0.995 against a required 0.5. The reported secondary closed-form correlations, 0.55 and 0.81, were not scored and do not change Adam's verdict. Ratios of arm medians are different statistics: observed median lag is 0.094 for both optimizers, whereas predicted integrated medians are 0.080 and 0.090 and closed-form medians are 0.089 and 0.088.

SGD's per-run integrated ratios have 10th–90th percentiles 1.04–1.06; every predicted run is within 10%, with median absolute lag error 0.005. Adam's corresponding ratios span 0.74–1.75, only 22% are within 10%, and median absolute error is 0.023. Adam's observed lag spread, 0.079–0.106, is narrower than its predicted spread, 0.054–0.129. Thus the test predicts typical Adam lag but not which runs lag more. The registered Adam test of per-run ordering failed (Spearman 0.25). Post hoc, the failure traces to when the preconditioner was measured: freezing it at the observed crossing instead of at half the switch scale raises the per-run Spearman to 0.99, with every run within 10%, showing that the rule-point preconditioner was stale by the crossing. A pre-crossing rule chosen after seeing these runs, taking the preconditioner at the first step the output scale reaches the occupied branch's switch, gives 0.95 (91% within 10%); we name it as the rule for a future registered test.

V.8 Interpretation of the earlier SGD own-sample test

With SGD, the registered own-sample predictions pass on the runs that crossed (30 of 40 at each a): each run's own threshold predicts its crossing better than the population threshold, and the rank correlation between crossing and own threshold is 0.67 and 0.68 (registered bound 0.6). When the 10 non-crossing runs at each a are counted at their final scale, the first prediction still holds, but the rank correlation falls to 0.50 and 0.51 and fails the bound. We therefore report the SGD result as a partial pass: it holds for crossers only. The census retains the original registered scoring population.

W REGISTERED RAMPS AND LEARNING-RATE INVARIANCE

The ramp experiment fixes an exponential scale path after a 4,000-step warm-up from half the population switch scale. Each setting has six growth rates and 40 fresh seeds per rate. The output bias remains a trained coordinate. All 1,440 runs cross and none changes winding. Own thresholds were frozen before comparisons, but the registered reference is each sample's global threshold rather than the switch of the forced branch. These verdicts are mixed. In a post hoc analysis against each forced

Table 41: Registered ramp results. The columns are the through-origin slope, signed Spearman correlation and slowest-rate median lag. P and F are the registered pass and fail verdicts.

a	Winding	Optimizer	Slope	Rank	Slowest lag
1.30	-1	Adam	-7.12 F	1.00 P	-0.0000 P
1.30	-1	SGD	0.78 P	1.00 P	+0.0009 P
1.30	0	Adam	9.10 F	1.00 P	-0.0012 P
1.30	0	SGD	1.30 F	1.00 P	-0.0014 P
1.50	0	Adam	-9.99 F	0.71 F	-0.0250 F
1.50	0	SGD	-0.10 F	1.00 P	-0.0233 F

branch's own-sample switch, SGD cell-median lags agree within 6% in the four slowest cells. They agree within 30% across the measured range. The negative-constant winding crosses early, as its sign predicts. This changes no registered verdict.

Adam fails even under that post hoc reference. Warm-up at a stationary point reduces its square-root second-moment estimate by 8–88 times relative to free training. Its frozen-preconditioner relaxation time is then 0.17–0.19 steps, outside the small-step, stationary-preconditioner regime. A follow-up ramp from initialization found no admissible growth rate under its calibration rules, so it was never registered. The Adam form of the law has support in free training, where growth is not imposed. In free training both output growth and hidden relaxation scale with the learning rate, leaving their

Table 42: Registered free-Adam learning-rate invariance test. Each cell has 39 crossings of 40 runs. Tolerances compare the lower-rate medians with the base rate, not uncertainty intervals. Both activation settings pass.

a	$\eta = .01$	.005	.0025	Tolerance
1.30	0.0296	0.0304	0.0308	± 0.0100
1.50	0.0619	0.0625	0.0647	± 0.0155

ratio unchanged. The model therefore predicts a lag that does not vanish in the gradient-flow limit. The finite learning-rate test supports this two-timescale interpretation; it does not itself compute that limit.

X WIDTH TWO: POST HOC BRANCHES AND THE VALIDITY CONDITION

All analyses here are post hoc. The full-speed failure, unresolved follow-ups, slowed comparison failure and no-gating failure in Appendix N retain their registered verdicts.

Conditional minima on different objectives. On the population objective the global minimizer at the same scale is placed and every stuck state lies above it (by 10^{-5}–10^{-4}); on the finite training sets the placed and stuck minima are nearly degenerate, and the stuck one is often lower. The population comparison covers the held scales with a validated placed minimum, $R_2 = 0.003, 0.01$. All 190 corresponding states lie above it; no validated population comparison is claimed for the remaining held levels.

All 265 stationary unplaced sine endpoints reduce to a duplicated unit with equal weight split. Weight transfer between identical units gives an exactly flat direction; the transverse Hessian is positive definite. These are conditional Morse–Bott local minima, not isolated strict minima. On each run's own sample, a distinct placed minimum was found for 58 endpoints; the stuck state has

lower loss in 37 of those cases. None of these sample searches certifies global optimality. Tanh has no identified local minimum among its 16 stationary-by-criterion endpoints: further polishing yields saddles or undetermined states.

Tracking along the output-weight path. A width-2 branch is continued in the hidden coordinates at the run's measured output-weight vector, following both its growth and rotation. Its switch is a per-run, per-path quantity computed after observing the trajectory. Three pilot runs preceded the choice of path continuation over continuation along a fixed ray. The primary path analysis uses every eligible run.

A branch is defined for 191 of 230 slowed runs. Most of the undefined cases cross during initial relaxation into an already placed basin, before a quasi-static description applies. Among 165 slowed branch-defined runs in the width-1 timescale range, 77% lie within $\max(0.01, 0.25|\text{prediction}|)$, with median $|\log(s_c/s^*)| = 0.003$. Their predicted lags are below the tolerance floor, so they test only the near-zero-lag limit. Systematic slightly early crossings in some slowed arms remain unexplained.

At full speed, the median crossing ratio χ is 0.99. Most runs cross far from a branch, within about a relaxation time. Three full-speed runs have χ within the width-1 range; their measurable lags have observed/predicted ratio 1.04. Their state distance exceeds the declared on-branch tolerance, so they are not classified as on-branch runs. This limited agreement does not validate the law at full-speed width-2 training generally.

The post hoc validity condition. The data support tracking for $\chi \lesssim 0.06$ after branch relaxation. Among width-2 branch-defined runs, agreement drops sharply outside that range. Early relaxation crossers can violate tracking even at small χ. A proposed prospective early-scale rule required a final branch classification before the earliest crossing, but no admissible scale existed. Nothing was registered under that design.

A separate predictor computed only from the initial state combined fixed-scale relaxation with continuation along the initial output-weight ray. Its post hoc gate accuracy was 81%, below the required 95%. The late-class median absolute log error is 0.014, but the gate fails on identifying early crossers. Of the early crossers, 42 of 96 are missed. Most missed cases follow a branch whose switch is driven by growth and rotation along the actual output-weight path. No prospective test was registered, and the late-class accuracy is not prospective evidence.

Y Additional activation and input-dimension tests

Y.1 Practical non-monotone activations

GELU, SiLU and Mish each have a validated conditional placement switch, without a global interval certificate. The small-scale class-mean maximum is not attained, so the criterion's attainment hypothesis cannot simply be assumed. The limiting ramp analysis predicts a switch for GELU; it is undetermined for SiLU and Mish. Training uses Adam at learning rate 0.01 on 400 balanced points

Table 43: Validated practical-activation geometry and conditional switch. These are numerical values and brackets, not certified global enclosures.

Activation	$\widehat{G}$	Output-scale bracket	Branch switch
GELU	0.037547	$[6.6117, 6.6714]$	6.6456
SiLU	0.050912	$[3.6190, 3.6517]$	3.6392
Mish	0.054772	$[3.1908, 3.2197]$	3.2147

for 32,000 steps. The registered primary test has 40 seeds per activation and remains unresolved because too few runs cross. A secondary 200-seed arm was registered before data collection and fails both the above-threshold and lag comparisons. Post hoc, a tracked-branch switch is defined for 117 GELU crossers. Their median χ is 0.0137, within the primary setting's range. Their median lag is $+0.0017$ against prediction $+0.0021$, and 96% cross within 1% of the branch switch. The

Table 44: Registered practical-activation results. Primary counts fail the minimum-crossing gate; secondary threshold and lag predictions fail for every activation. Lags use the population threshold.

Activation	Primary crossed	Secondary crossed	Above fraction	Observed lag	Predicted
GELU	23/40	132/200	0.477	-0.0183	+0.0023
SiLU	25/40	135/200	0.622	+0.0353	-0.0324
Mish	25/40	135/200	0.704	+0.0658	-0.0393

other early GELU crossers lie on a tail-placed ramp branch with no continued switch. SiLU and Mish have median ratios 0.2242 and 0.2556; their branch-referenced lags have the wrong sign for the linear law. This is consistent with the measured validity boundary, not a registered confirmation of it.

The proposed GELU early-branch rule could not satisfy its pre-registration design conditions: most runs were already initialized above the earliest crossing scale, and the initial basin rarely matched the crossing branch. The prospective follow-up was therefore never registered. In these activations the branch-conditioned account is explanatory only.

Y.2 The band task with additional input coordinates

A single sine unit receives the original signal coordinate and independent uniform noise coordinates. Its output is $w_2 f_a(w \cdot x + b_1) + b_2$. The continuous-support gap widens every projected window by $2\|w_{\text{noise}}\|_1$. At zero noise weight it agrees exactly with the original gap.

The population loss is an average $\mathbb{E}_\xi L_1(w_1, b_1 + \xi, b_2)$, with label-independent $\xi = w_{\text{noise}} \cdot x_{\text{noise}}$. It is at least the width-1 minimum, attained at zero noise weight. The certified width-1 threshold therefore carries over. A separate population numerical check is validated, not certified. No search candidate improves on the width-1 minimum.

The design was fixed in code before higher-dimensional training, but the registration document followed a disclosed pilot that recorded crossing counts and times without scales. The pilot changed the budget and added a secondary arm. Training uses Adam at learning rate 0.01 on 400 points for 64,000 steps. Primary cells have 40 seeds. The registered secondary arm has 120 seeds per setting.

Table 45: Registered band-task results. P1 tests the fraction crossing at or above the certified width-1 threshold. P2a compares the population-referenced lag with the width-1 range. P2b compares lag above the signal-sample own threshold with the frozen law.

d	a	Crossed	P1	P2a	P2b: observed / predicted
2	1.30	36/40	0.972, pass	0.182, fail	0.147 / 0.024, fail
2	1.50	36/40	0.972, pass	0.207, fail	0.187 / 0.049, fail
4	1.30	29/40	unresolved	unresolved	unresolved
4	1.50	29/40	unresolved	unresolved	unresolved

Its above-threshold test passes in every cell, while its population-lag-range test fails in every cell. It does not replace the unresolved primary result at the higher dimension. Non-crossing runs end noise-dominated; no claim is made that a longer budget resolves them.

Post hoc continuation on each run's own higher-dimensional sample separates a static threshold shift from dynamic lag. The finite-sample minimizer retains a small noise weight, widening the continuous-support windows. Its own threshold rises by about 12% in the lower dimension and 40% in the higher dimension. Against that threshold the lag agrees with the law in every cell, with observed/predicted ratios 0.84–1.10. This branch-conditioned decomposition was not registered and is not a law in dimension or sample size. The population threshold itself does not move.

Y.3 Populations behind the prospective scoreboard

Figure 16 includes registered value or range predictions, not every prediction in the census. The new-activation-value test has 80 runs per value; SGD has 40; the learning-rate test has 40 per cell.

Table 46: Post hoc higher-dimensional branch comparison. Entries are median observed and predicted lags with their ratio. The population-reference failures above retain their verdicts.

Population	$a = 1.30$	$a = 1.50$
$d = 2$, primary	0.025 / 0.024 (1.06)	0.054 / 0.049 (1.10)
$d = 4$, primary	0.019 / 0.017 (1.09)	0.038 / 0.036 (1.07)
$d = 4$, pooled	0.014 / 0.017 (0.84)	0.033 / 0.034 (0.96)

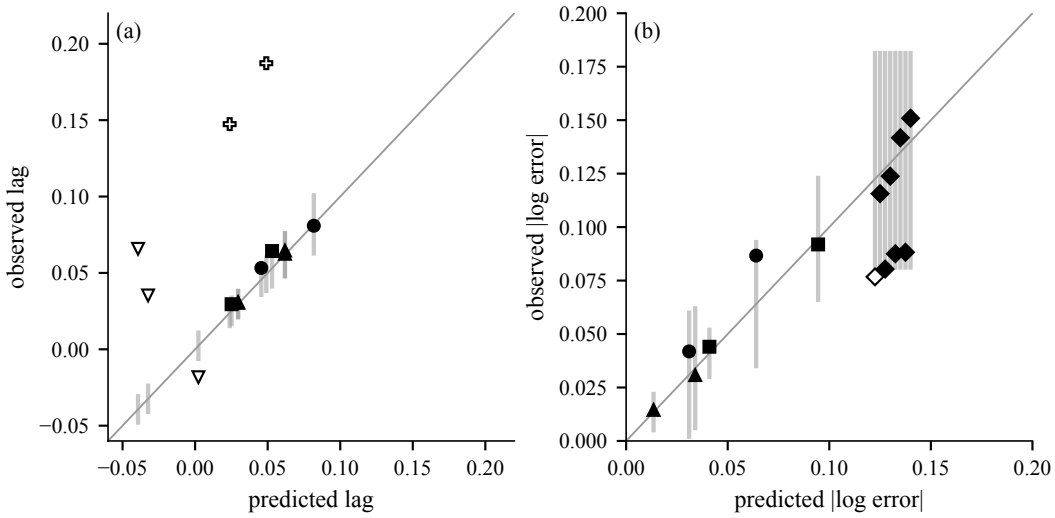


Figure 16: Earlier registered value predictions succeed in the primary sine setting and fail for the practical activations and the band task. Filled markers lie within the registered tolerance or range (grey bars); open markers lie outside. Counts within are: fresh-sample timescale $2/2$, SGD $2/2$, learning-rate invariance $4/4$, practical activations $0/3$ (200 seeds each), single unit in $\mathbb{R}^2$ $0/2$, own-sample threshold $2/2$, early branch $2/2$, lag-corrected early branch $2/2$, and held-out windows $7/8$. The window test has 90 runs per setting; other populations are given in Appendix Y.3. Paired-comparison predictions, all passing, and forced-ramp slope tests are not point predictions and are omitted.

The own-threshold test has 60 per setting, and the held-out-window test has 90. The practical-activation secondary tests use 200 seeds; the band-task primary cells use 40. Their methods and verdicts are given in the corresponding appendices.

The plotted counts sum to 21 of 27 within tolerance or range. The outside points comprise the practical activations, the higher-dimensional band task, and the held-out geometry whose offset lies outside its registered range. Paired comparisons without a specified predicted value are omitted: the own-versus-population comparisons under Adam and SGD, the new-activation-value comparisons, other own-threshold criteria, and the geometry predictor versus its baselines. All these paired comparisons pass. The forced-ramp tests concern slopes and rank correlations and are not plotted as point predictions.