WHAT LEARNING-RATE SUSCEPTIBILITY MEASURES: COORDINATES AND TRAINING ROUTES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-rate susceptibility is often interpreted as an intrinsic indicator of training instability, although its value can mix two effects: the coordinate used to perturb the learning rate and changes in the optimization route. We formalize this distinction. Finite predictor-space KL is invariant under coordinate relabeling for a fixed physical intervention, whereas its normalized infinitesimal coefficient rescales with the intervention coordinate. For ReLU–SGD, we give a two-update construction in which bounded smooth-branch Jacobians coexist with an initialization-averaged response $c|\delta| + O(\delta^2)$ because an $O(|\delta|)$ set of initializations crosses a moving training boundary. We then test the route mechanism in a frozen SVHN/TinyResNet experiment with 20 fresh successor trajectories. Replaying the base trajectory's activation gates and max-pool indices reduces median conditional-quadratic forecast discrepancy in every trajectory by 93.62–99.67%; 15/20 satisfy the predeclared event (one-sided exact $p = .0207$; 95% Clopper–Pearson interval $[.509, .913]$). Matched controls support route mediation rather than automatic-differentiation dominance, and absolute prediction changes remain small. Thus susceptibility is a coordinate- and protocol-dependent response diagnostic; by itself it does not predict accuracy, divergence, or learning-rate utility.

1 INTRODUCTION

Changing a learning rate changes both the size of an update and, potentially, the activation and pooling routes followed by a nonsmooth network. A susceptibility score compresses the resulting predictor change into one number. Its interpretation depends on two choices that are often left implicit: how the rate is perturbed and whether the measured response includes changes in the training route. We ask what can be inferred from the score once those choices are made explicit.

We compare trajectories that share initialization and stochastic inputs but use different rates. For a specified pair of rates, their prediction-space KL measures the finite response to that physical intervention. A quotient formed by dividing that KL by a squared rate increment depends on the units of the increment. This matters when a peak in susceptibility is read as a training boundary or used to choose a rate. Curvature, divergence, final validation loss, and predictor motion can occur near one another (Cohen et al., 2021; Lewkowycz et al., 2020; Damian et al., 2023); they describe distinct aspects of training.

The analysis first separates the physical intervention from its coordinate description. A coordinate change leaves KL between fixed predictor pairs alone, but rescales a local quadratic coefficient and can change its ranking across rates. An exact signed-loss identity then shows how the linear, label-aligned term supplies the orientation of prediction movement. The second distinction is between a derivative along one smooth training route and an average over routes. In a two-update ReLU–SGD example, a moving boundary sends an $O(|\delta|)$ fraction of initializations across a finite predictor jump. The averaged KL is therefore $c|\delta| + O(\delta^2)$ even though the derivative on almost every fixed branch is bounded.

We then intervene on the route itself. After identifying and repairing a route-recording error, we froze a new 20-trajectory SVHN/TinyResNet protocol. Replaying the base trajectory's ReLU and pooling decisions reduces the median conditional-quadratic forecast discrepancy in every trajectory; 15 meet the complete predeclared event. The paired controls point to route changes as a source of the

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.

forecast error, while a logit secant has slightly lower error than conditional automatic differentiation after replay. Absolute prediction changes are small.

Together, the theory and experiments give a working interpretation of susceptibility: report the coordinate and physical endpoints, then ask whether the measured response includes changes in the training route. The route test provides evidence over a short horizon in one protocol; the earlier forecasting and rate-selection studies show how the same measurements behave in longer-horizon decisions.

2 Related work

Training dynamics and diagnostics. Work on the edge of stability, catapult dynamics, and bifurcations relates learning rate to the evolution of the training state (Cohen et al., 2021; Lewkowycz et al., 2020; Damian et al., 2023; Kalra et al., 2023; Ghosh et al., 2025). Stochastic Hessian products and momentum affect that evolution (Wu & Su, 2023; Andreyev & Beneventano, 2024; Andreyev et al., 2026); spectral alignment and early telemetry offer different ways to forecast outcomes (Qiu et al., 2025; Naik et al., 2026). These methods concern training outcomes; the present experiments measure predictor response to a specified rate intervention.

Hyperparameter derivatives and Fisher geometry. Forward differentiation through an optimizer yields the forced tangent recursion used below (Franceschi et al., 2017). Related hypergradient methods optimize validation objectives (Maclaurin et al., 2015; Baydin et al., 2018; Franceschi et al., 2018; Lorraine et al., 2020). Their signed derivative, for example $\partial_\eta L_{\mathrm{val}}(\theta_T) = \nabla L_{\mathrm{val}}(\theta_T)^\top s_T$, answers a different question from a squared magnitude of predictor motion. Long-horizon differentiation addresses memory and gradient degradation, while a short-horizon objective may favor a rate that performs poorly later (Micaelli & Storkey, 2021; Wu et al., 2018). Repeated Jacobians can amplify unrolled derivatives (Metz et al., 2022); differentiating an expectation through a nonsmooth map may also require a boundary contribution (Lee et al., 2018). Our two-update example isolates that boundary contribution for a coupled rate intervention. The Fisher coordinate law and invariant line element use standard information geometry (Amari & Nagaoka, 2000; Martens, 2020; Karakida et al., 2019).

Rate choice and function-space motion. Range tests, cyclical schedules, learning-rate-free methods, and curvature-aware tuners use training measurements to make rate decisions (Smith, 2017; 2018; Defazio & Mishchenko, 2023; Cutkosky et al., 2023; Roulet et al., 2024). Function-space learning rates measure output change from an optimizer update (Milsom et al., 2025), and fitted loss curves forecast outcomes under new schedules (Luo et al., 2025). Our intervention instead changes the rate along a finite trajectory. In the coordinate comparison, the physical predictor trajectories are held fixed, unlike a change of parameterization in sharpness analyses (Dinh et al., 2017). The disagreement bound below applies Pinsker's inequality to those finite predictors (Reid & Williamson, 2011).

3 Response dynamics and coordinate geometry

Throughout this section, two rate trajectories share their initial state and training randomness. For $\theta_{k+1} = G_{\eta,k}(\theta_k)$, differentiation gives

$$s_{k+1} = A_k s_k + b_k, \qquad A_k = D_\theta G_{\eta,k}(\theta_k), \quad b_k = \partial_\eta G_{\eta,k}(\theta_k). \tag{1}$$

For SGD, $A_k = I - \eta H_k$ and $b_k = -g_k$. Stateful optimizers use an augmented state $z_k = (\theta_k, v_k, \dots)$ and the parameter component of $\partial_\eta z_k$. Theorem 2 expands this recursion into a sum of forced tangent products. The finite-horizon upper bound grows as contraction weakens. Cancellation among forcing terms and projection into predictor space determine how much of that potential appears in the measured response.

Fix a probe distribution Q. For a regular predictor, let

$$F_Q(\theta) = \mathbb{E}_{x\sim Q,\, y\sim p_\theta(\cdot|x)}[\nabla_\theta \log p_\theta(y \mid x) \nabla_\theta \log p_\theta(y \mid x)^\top], \qquad q_k^\eta = \tfrac{1}{2} s_k^\top F_Q(\theta_k) s_k. \tag{2}$$

The factor $1/2$ matches the quadratic KL coefficient. We distinguish the average and maximum

$$\mathcal{S}_T^{F,\eta} = \frac{1}{T}\sum_{k=1}^{T} q_k^\eta, \qquad \mathcal{S}_{\max}^{F,\eta} = \max_{1\le k\le T} q_k^\eta. \tag{3}$$

We write $F = F_Q$ when the probe is fixed. These statistics differ from the terminal Euclidean sensitivity in Proposition 6. All divergences use natural logarithms, and prediction scores average per-input divergence over Q. We use either of two declared KL conventions:

$$D_{\to}(p,q) = D_{\mathrm{KL}}(p\|q), \qquad D_{\mathrm{sym}}(p,q) = \tfrac{1}{2}\{D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)\}. \tag{4}$$

Proposition 1 (Local shadow expansion)**.** *Suppose the predictor has positive finite-class probabilities and $\log p_\theta(y \mid x)$ is twice continuously differentiable near $\theta_k(\eta)$. Assume that the parameter derivatives through order two of both conditional directional KL integrands admit Q-integrable envelopes on that neighborhood, and that F_Q is finite and continuous. If $\theta_k(\eta+\delta)-\theta_k(\eta) = \delta s_k + o(\delta)$, then for either convention in Equation* (4)*,*

$$\frac{\mathbb{E}_Q D_\diamond(p_{\theta_k(\eta)}(\cdot \mid x), p_{\theta_k(\eta+\delta)}(\cdot \mid x))}{\delta^2} = q_k^\eta + o(1). \tag{5}$$

On a fixed finite probe, smooth finite-class softmax predictors satisfy these local conditions. A finite rate change in a ReLU network can cross an activation boundary. For the image experiments we therefore report the finite statistic

$$\widehat{\mathcal{S}}_{\max}^{\mathrm{KL},\eta}(\eta) = \frac{R_{\max}(\eta, \eta+\delta)}{\delta^2}, \qquad R_{\max}(\eta,\eta') = \max_{1\le k\le T_{\mathrm{probe}}} \mathbb{E}_Q D_\diamond(p_{k,\eta}, p_{k,\eta'}). \tag{6}$$

Here $p_{k,\eta} = p_{\theta_k(\eta)}(\cdot \mid x)$. The maximum is outside the input expectation.

Proposition 2 (Coordinate covariance)**.** *Let $u = h(\eta)$ be an increasing C^1 diffeomorphism between open intervals. For the same trajectory family,*

$$\mathcal{S}_T^{F,u}(h(\eta)) = \left(\frac{d\eta}{du}\right)^2 \mathcal{S}_T^{F,\eta}(\eta). \tag{7}$$

The same law holds for $\mathcal{S}_{\max}^F$. For any two distinct rates with positive coefficients, smooth increasing coordinates exist that induce either strict pointwise ranking.

In particular, for $\eta > 0$,

$$\mathcal{S}_T^{F,\log\eta} = \eta^2 \mathcal{S}_T^{F,\eta}. \tag{8}$$

The Jacobian is independent of the checkpoint, so the same multiplication applies to the maximum.

Proposition 3 (Invariant interval length)**.** *Let h be an increasing C^1 diffeomorphism as in Proposition 2. For fixed physical endpoints $\eta_a < \eta_b$ in its domain, if the integral is finite, then*

$$\mathcal{L}_T([\eta_a,\eta_b]) = \int_{\eta_a}^{\eta_b} \sqrt{\mathcal{S}_T^{F,\eta}(\eta)}\, d\eta = \int_{h(\eta_a)}^{h(\eta_b)} \sqrt{\mathcal{S}_T^{F,u}(u)}\, du. \tag{9}$$

The identity also holds using the maximum-based coefficient in both integrals.

Zeros of the coefficient do not affect the identity. On an interval where a continuous coefficient is positive, choosing $u(\eta) = \int^\eta \sqrt{\mathcal{S}_T^{F,\eta}(r)}\, dr$ makes that coefficient constant. A pointwise peak consequently has meaning only relative to a stated rate coordinate; the response accumulated across fixed physical endpoints has the invariant form above.

Exact finite coordinate comparison. For positive endpoints η and $(1+\epsilon)\eta$, where $\epsilon > -1$ and $\epsilon \neq 0$,

$$\widetilde{\mathcal{S}}_{\max}^{\mathrm{KL},\log\eta} := \eta^2 \widehat{\mathcal{S}}_{\max}^{\mathrm{KL},\eta} = \frac{R_{\max}}{\epsilon^2}, \qquad \widehat{\mathcal{S}}_{\max}^{\mathrm{KL},\Delta\log\eta} = \frac{R_{\max}}{[\log(1+\epsilon)]^2} = \frac{\epsilon^2}{[\log(1+\epsilon)]^2}\widetilde{\mathcal{S}}_{\max}^{\mathrm{KL},\log\eta}. \tag{10}$$

For fixed ϵ, the re-expressed coefficient, exact finite-log quotient, and raw $R_{\max}$ have identical rankings by algebra on the finite predictor pairs. For $\eta_i < \eta_j$ and positive additive scores S_i, S_j, a strict reversal occurs exactly when $1 < S_i/S_j < (\eta_j/\eta_i)^2$. Multiplication by increasing η^2 cannot move a unique maximizer downward. Whether a shift occurs is empirical; its upward direction is forced.

4 FINITE INTERVENTIONS AND INTERPRETABLE RESPONSE

A finite response requires a rule for choosing the alternative rate. At checkpoint k, fix an empirical or population probe law Q and a shared law for training randomness. Let K_η choose alternative rates independently of that randomness, and define

$$R_{K,k}(\eta) = \mathbb{E}_{\omega,\, \eta' \sim K_\eta,\, x \sim Q} D_\diamond(p_{k,\eta,\omega}(\cdot \mid x), p_{k,\eta',\omega}(\cdot \mid x)). \tag{11}$$

Assume the predictor and kernel maps are measurable. Under $u = h(\eta)$, transporting the law as $K_u^h = h_\# K_{h^{-1}(u)}$ leaves $R_{K,k}$ unchanged because it samples the same physical pairs with the same probabilities. Drawing new additive noise after changing coordinates generally samples different pairs. The saved 5% shadows take K_η to be a point mass at 1.05η; their raw KL is a finite response under that intervention.

Proposition 4 (Signed validation response and outer curvature)**.** *For logits $f, \widetilde{f} \in \mathbb{R}^C$, let $p = \mathrm{softmax}(f)$, $\widetilde{p} = \mathrm{softmax}(\widetilde{f})$, and $\ell_y(f) = -\log p_y$. Then*

$$\ell_y(\widetilde{f}) - \ell_y(f) = (p - e_y)^\top (\widetilde{f} - f) + D_{\mathrm{KL}}(p \| \widetilde{p}). \tag{12}$$

For a labeled validation set of size n, suppose trajectory logits $f_i(a)$ are twice differentiable in a schedule coordinate $a \in \mathbb{R}^d$. At a_0, set $p_i = \mathrm{softmax}(f_i(a_0))$, $J_i = D_a f_i(a_0)$, $F_i = \mathrm{diag}(p_i) - p_i p_i^\top$, and $H_{ic} = \nabla_a^2 f_{ic}(a_0)$. The mean cross-entropy L_V satisfies

$$\nabla_a L_V(a_0) = \frac{1}{n} \sum_i J_i^\top (p_i - e_{y_i}), \tag{13}$$

$$\nabla_a^2 L_V(a_0) = \underbrace{\frac{1}{n} \sum_i J_i^\top F_i J_i}_{G_V} + \underbrace{\frac{1}{n} \sum_{i,c} (p_{ic} - e_{y_i,c}) H_{ic}}_{C_V}. \tag{14}$$

The first identity uses KL from the base prediction to the perturbed prediction. Its linear term can have either sign, while the KL remainder is nonnegative. In the Hessian identity, G_V is the pullback Fisher or generalized Gauss–Newton term, and C_V records how the trajectory curves through logit space. The signed hypergradient measures label-aligned motion, and full outer curvature includes the trajectory term. Appendix B.1 proves these softmax and chain-rule identities and gives a stable two-step SGD example with $G_V > 0$ but $G_V + C_V < 0$. The Hessian and C_V are coordinate-specific: nonlinear reparameterization adds gradient-dependent terms away from stationarity, whereas G_V transforms as a metric.

4.1 A BOUNDARY MECHANISM FOR NONQUADRATIC MEAN RESPONSE

The local KL expansion holds along a regular training branch. Averaging over initializations requires control near boundaries between branches. Although a ReLU forward map is continuous, its gradient update can jump at an activation boundary. A probe on another input can detect that jump.

Proposition 5 (Two-update ReLU-SGD boundary response)**.** *Let σ be logistic, $w_* = \log 3$, and $\eta_* = 4\log(3)/3$. Let W have a bounded nonnegative density f, integrating to one and supported on a compact subset of $(1, 6/5)$, with f continuously differentiable near w_* and $f(w_*) > 0$. Train logits $(0, \mathrm{ReLU}(wx))$ on $(x, y) = (1, 0)$ with softmax cross-entropy $\ell(w) = \log(1 + \exp(\mathrm{ReLU}(w)))$, for two plain SGD updates at the same rate, from $w_0 = W$, with $\mathrm{ReLU}'(0) = 0$. Evaluate the class-one probability $p_\eta(W)$ at the fixed probe $x = -1$. Both rates share W. Put*

$$T(w) = w/\sigma(w), \quad \rho_T(t) = \frac{f(T^{-1}(t))}{T'(T^{-1}(t))}, \quad J(t) = \frac{t}{4}\left[\sigma(t/2) - \frac{1}{2}\right].$$

As $h \to 0$ with $\eta_* + h \in (7/5, 3/2)$, using the half-sum symmetric KL of Equation (4),

$$R(h) := \mathbb{E}_W D_{\mathrm{sym}}(\mathrm{Ber}(p_{\eta_*}(W)), \mathrm{Ber}(p_{\eta_*+h}(W))) = \rho_T(\eta_*) J(\eta_*) |h| + O(h^2). \tag{15}$$

For almost every fixed W, the same divergence divided by h^2 converges to a local Fisher coefficient $q(W) \in [0, 1/2]$. Thus $\mathbb{E}q(W) < \infty$, but $R(h)/h^2 \to \infty$.

Writing $u = w - \eta\sigma(w)$, the second state is $u - \eta\sigma(u)$ for $\eta < T(w)$ and u for $\eta > T(w)$. At the threshold its one-sided limits are $-T(w)/2$ and zero. The negative-input probe sees this jump; the training-input prediction is $1/2$ on both sides. Every smooth-branch state Jacobian lies in $[5/8, 1]$, but the global SGD map is discontinuous at zero. Appendix B.2 proves the result and gives the expected-loss boundary term, which can reverse the mean pathwise hypergradient's sign.

The cusp describes an initialization population. In any fixed finite sample, almost surely the nearest threshold is some positive distance from η_*, and smaller interventions recover quadratic empirical behavior. For an independent bounded Z and small $a > 0$, the same construction gives a leading term $\rho_T(\eta_*)J(\eta_*)a\mathbb{E}|Z|$ under a random intervention aZ. A forecast based only on its second moment misses that term. The construction identifies a route-boundary mechanism for this departure from quadratic response.

4.2 Finite certificates and local rankings

Theorem 1 (Finite prediction-disagreement certificate)**.** *Let p_x, q_x be measurable distributions on a finite class set of size at least two, with a fixed deterministic tie-breaking rule. Write $c(x) = \arg\max_j p_x(j)$ and $\kappa(x) = p_x(c(x)) - \max_{j \neq c(x)} p_x(j)$, with $\kappa = 0$ at base ties. For either declared KL convention let $R = \mathbb{E}_Q D_\diamond(p_x, q_x)$. Then for every $\gamma > 0$,*

$$Q\{\arg\max q_x \neq c(x)\} \leq \min\left\{1,\ Q\{\kappa \leq \gamma\} + \frac{2R}{\gamma^2}\right\}. \tag{16}$$

If per-example divergences are available, a second bound is $\mathbb{E}_Q \min\{1, 2D_\diamond(p_x, q_x)/\kappa(x)^2\}$, with integrand one at base ties.

A class change requires $\|p_x - q_x\|_1 \geq \kappa(x)$; Pinsker's inequality then gives $D_\diamond \geq \kappa^2/2$. This finite bound applies across nonsmooth paths and also controls the absolute change in zero-one error on the same input distribution. Averaging under K_η and the shared training law preserves the bound and its coordinate invariance. An empirical Q yields an empirical-probe guarantee; sampling assumptions extend it to a population. Appendix C distinguishes worst-checkpoint probability from the probability that an input changes class at any checkpoint.

Finite quotients and derivative rankings. Let $\phi_{ik}(r) = \mathbb{E}_Q D_\diamond(p_{k,\eta_i}, p_{k,\eta_i+r})$. For $\delta_i \neq 0$, if ϕ_{ik} is C^3 on a neighborhood of the interval between 0 and δ_i, has $\phi_{ik}(0) = \phi'_{ik}(0) = 0$, $\phi''_{ik}(0) = 2q_{ik}$, and $|\phi'''_{ik}| \leq M_{ik}$ there, Taylor's theorem gives

$$|\phi_{ik}(\delta_i)/\delta_i^2 - q_{ik}| \leq e_{ik} := M_{ik}|\delta_i|/6. \tag{17}$$

Hence maximum scores have error at most $E_i = \max_k e_{ik}$, and an observed gap exceeding $E_i + E_j$ certifies the local ordering in the declared coordinate. Weighted averages use the weighted error bound. Appendix D supplies sufficient log-probability and trajectory derivative bounds. Symmetric KL between the same one-sided endpoints still has first-order quotient error in general. The archived image records support the finite ranking identity at their observed endpoints; a derivative-error certificate requires additional trajectory constants.

5 Experiments

We begin with the route intervention that tests the proposed mechanism. We then examine how finite scores change under coordinate normalization and how well a finite quotient matches its local coefficient in a smooth network. Finally, we report the earlier forecasting and selection tests that limit the use of these measurements. The image datasets are MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), Kuzushiji-MNIST (Clanuwat et al., 2018), and SVHN (Netzer et al., 2011). Coordinate comparisons keep the physical endpoints and KL convention fixed; trajectory interventions pair initialization, optimizer state, and stochastic inputs. Appendix G gives data lineage and full settings.

5.1 Prospective training-route intervention

The route recorder observes ReLU gates and max-pool indices during the gradient forward itself. After a separate instrument check, we froze a new set of 20 SVHN/TinyResNet CPU trajectories at

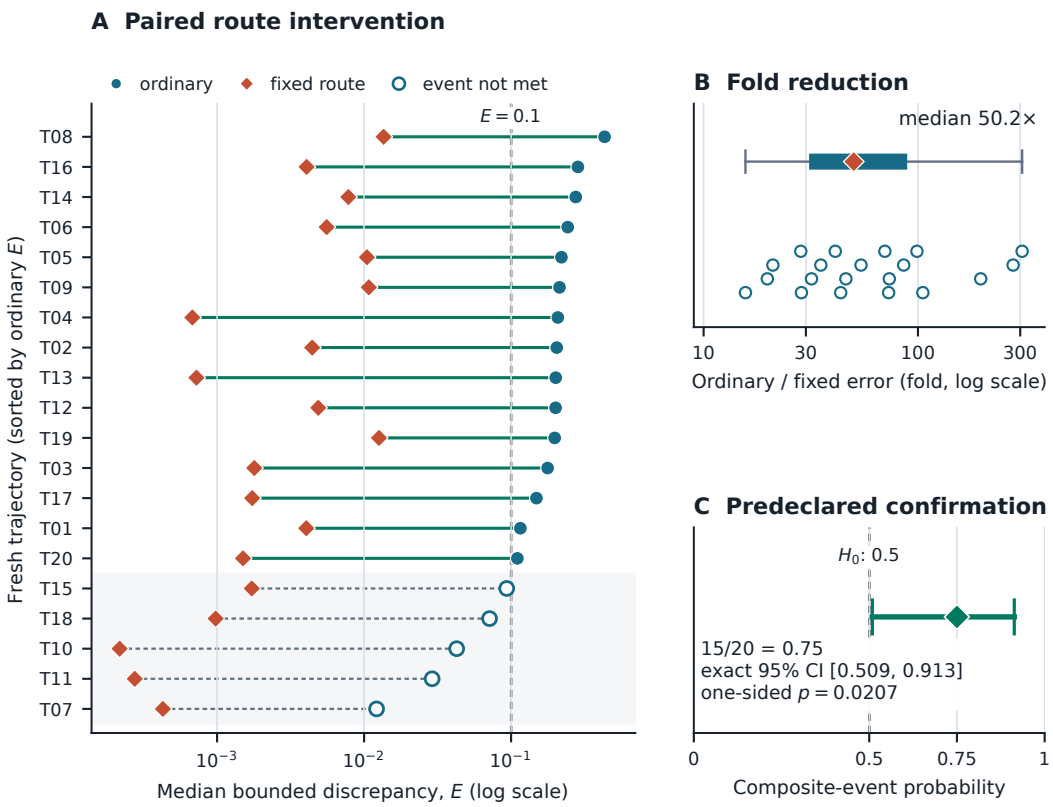


Figure 1: Prospective route confirmation on 20 fresh SVHN/TinyResNet CPU trajectories. (A) Paired trajectory medians of bounded discrepancy $E = |a - b|/\max(a, b, 10^{-10})$, sorted by ordinary E; shapes distinguish training conditions, open circles and the shaded rows mark the five event failures, and the dashed line is $E = .1$. (B) All 20 ordinary/fixed-route error ratios on a log scale, with range, interquartile range, and median. (C) The 15/20 event proportion and exact 95% Clopper–Pearson interval against $H_0 : p = .5$. All trajectories pass checks, resolve 20/20 cases, and exceed 50% reduction; the five non-events have ordinary $E < .1$.

update 200, each with its own hash-derived initialization and minibatch stream. Ordinary perturbed training follows whatever activation and pooling routes it encounters. In the intervention, the same perturbed updates instead use the base trajectory's recorded routes; prediction at evaluation still uses natural inference.

The predeclared event has four parts: all correctness checks pass, at least 16 of 20 paired float64 cases resolve across $H = 2, 5$, ordinary median bounded discrepancy E is at least .1, and route replay cuts that paired median by at least half. With $n = 20$, the fixed rejection boundary for $H_0 : p \leq .5$ at nominal $\alpha = .025$ is 15 events.

All trajectories pass the checks and resolve 20/20 primary cases. Fifteen meet the event ($\hat{p} = .75$, one-sided exact $p = .0206947$, exact 95% Clopper–Pearson interval $[.508954, .913429]$). Replay reduces median E by 93.62–99.67% in every trajectory; the median relative reduction is 97.99%, and the median fold ratio is $50.2\times$. Across trajectories, median ordinary E is .199305 and median replay E is .002917. Each of the five non-events has ordinary $E < .1$, despite a large reduction.

The controls help locate the error. Under replay, matching-sign logit secants have median $E = .004198$, compared with .005959 for conditional automatic differentiation (Table 1). Both methods become more accurate when the training route is fixed, and the secant has the lower median under replay. Median absolute symmetric KL is 2.48×10^{-7}, placing the relative forecast improvement beside the small scale of predictor motion. The development profile and stopped predecessor motivated the test; the 20 new trajectories form its analysis roster. Appendices K and L document the two records.

Table 1: Matched forecast controls in the route confirmation. Each cell is the median across 20 trajectory-specific median bounded discrepancies, using the same 12 coarser float64 cases per method and condition; lower is better. These descriptive controls were not part of the binary confirmation test.

Forecast	Ordinary	Fixed routes
Conditional AD quadratic	0.124345	0.005959
Conditional logit linearization	0.124878	0.005292
Matching-sign logit secant	0.218905	0.004198
Fixed power 2	0.174705	0.006028
Fine-grid quadratic	0.271898	0.005982
Fixed power 1	0.790072	0.799995
Fixed power 0	0.957295	0.959999

5.2 Coordinate dependence in finite interventions

In an eight-seed SVHN development study, rates $.008$ and $.048$ were chosen inside a probe-loss tolerance band and then trained for 8,000 updates. The lower rate had better later loss, while additive-rate and log-rate normalizations gave opposing susceptibility contrasts (Table 7). The rates were close by the declared tolerance, but their high-minus-low probe-loss gap was $.00919$ (95% paired t interval $[.00445, .01392]$). The contrast illustrates coordinate dependence at these two selected rates. The residual probe-loss gap and the documented use of the official test split during development make the later official-test outcomes descriptive. Appendix G.1 gives the selection rule, the complete effects, the distinction between written criteria and runner flags, and the limits of the local protocol record.

The wider saved collection has 19 eligible seed curves and 276 complete-probe candidate rows. Every curve contains a ranking reversal and changes its maximizing rate under the coordinate conversion. In all, 639/1929 candidate pairs reverse (33.13% pooled); the equal-curve mean is 36.47%, with a block-preserving 95% interval of $[32.71\%, 40.40\%]$ conditional on this fixed roster (Appendix G.4). A separate post-hoc cohort changes the maximum in 17/24 curves and reverses 406/1872 pairs. Both comparisons use the same physical predictor pairs under two normalizations. In the post-hoc cohort, selecting the relative-response maximum has higher regret than selecting by probe loss (Appendix G.3).

5.3 Smooth-regime finite-shadow calibration

A smooth-network experiment checks the local approximation under conditions where its derivative can be computed. It uses a 62-parameter tanh MLP on two-moons data, three seeds, three rates $\{0.1, 0.5, 1.0\}$, 40 full-batch updates, and relative increments $\epsilon \in \{.005, .01, .025, .05\}$. Hessian-vector products propagate the exact forced tangent; logit Jacobian-vector products give q_k. All nine base runs and 36 ordinary SGD shadows complete without skipped or clipped updates, and every checkpoint is retained. Table 2 compares their finite quotients with the exact average and maximum coefficients.

Table 2: Absolute relative error (%) against exact Fisher coefficients, median over nine fixed seed/rate pairs. The columns distinguish KL convention and checkpoint aggregation. Results describe the fixed two-moons design.

Relative step	Forward mean	Symmetric mean	Forward maximum	Symmetric maximum
0.5%	0.549	0.594	0.341	0.440
1.0%	1.093	1.183	0.681	0.877
2.5%	2.698	2.918	1.690	2.174
5.0%	5.283	5.706	3.342	4.287

At a 5% increment, forward mean errors range from 2.785% to 6.128%, and symmetric mean errors from 3.727% to 6.586%. Errors decrease with the increment in this design. Symmetrizing KL at the same endpoints generally retains a first-order finite-quotient error.

For each seed, Q is the fixed 119-example validation empirical distribution. We evaluate Equation (16) over the frozen margin grid $\gamma \in \{.05, .1, .2, .3\}$. All 5,904 checkpoint/shadow/margin rows satisfy both KL versions of the inequality. Among 1,440 positive-step comparisons, 57 exhibit any disagreement; the maximum is $4/119 = 3.36\%$. The minimum-grid forward bound is below one in all 1,440 comparisons, with mean 3.785% and maximum 72.791%. Thus the certificate is nonvacuous throughout this empirical probe, with considerable variation in tightness. Appendix M gives full ranges and the protocol.

5.4 Forecasting and rate-selection tests

A 128-update branch pilot asked whether local quadratic calibration predicts finite response after training a ReLU network. It used two seeds in each of MNIST/SGD and SVHN/momentum, checkpoints at updates 200 and 1,000, three anchor rates, and matched absolute and relative interventions. All 248 physical branches completed, yet both regimes failed the specified forecast and refinement criteria. Recomputing the forecast on common inputs after the outcome was known left median discrepancies of 90.80% on MNIST and 93.21% on SVHN across the 48 anchor/law rows. The high discrepancy persists after the input sets are aligned. Unfitted constant and linear controls give different errors across regimes, pointing to variation in finite-response scaling. The complete endpoints and definitions are in Appendices H and I and Table 14.

The saved logits also check the signed-loss identity on 160 distinct base/target pairs; the largest per-example residual is 3.60×10^{-14}. Mean cross-entropy falls for 34/80 MNIST pairs and 46/80 SVHN pairs. The sign varies across these paired responses, as the linear term in Equation (12) predicts.

A separate selector study tested whether response and margin features improve decisions beyond rate and probe loss. Its model was fixed before evaluation on 24 held-out seed blocks across three datasets. All 442 candidate pairs completed, and every finite disagreement certificate equaled one. Held-out mean regret was .1543 for probe loss, .1660 for rate plus probe loss, and .1053 for the full model. The predeclared full-minus-rate-plus-probe contrast was $-.0607$, with a 95% block-bootstrap interval of $[-.1972, .0266]$ that crosses zero. The policies chose the same rate in 19/24 blocks, and one block contributed more than the net improvement. Under the planned transfer rule, the study stops at this stage. Table 5 and Appendices G.2–G.3 give the protocol and influence checks.

The earlier five-update route profile recorded routes in a separate forward pass, which could differ from the gradient forward. A subsequent 32-trajectory campaign stopped at its instrument check. The new 20-trajectory test uses the corrected recorder and its own analysis roster. The earlier records are documented in Appendices J and L.

6 Discussion

Susceptibility must be reported with the intervention that defines it. KL between fixed predictor pairs describes a finite response. Dividing by a squared coordinate increment produces a local coefficient whose value and ranking depend on that coordinate. The physical endpoints or intervention law, probe distribution, KL convention, and checkpoint aggregation specify what a reported score measures. The invariant interval length compares accumulated local response across fixed physical endpoints.

Route replay supplies the main empirical finding. After a 25-origin base-path check of the corrected recorder, base-route replay reduced conditional-quadratic forecast discrepancy in all 20 new trajectories over five updates; 15 met the full predeclared event. The remaining five had small ordinary discrepancy. Together with the matched controls, this pattern locates much of the forecast error in route changes. The small absolute KL scale places the relative improvement in the context of modest predictor motion. The experiment concerns exposed SVHN data, TinyResNet, CPU execution, and update-200 origins. The scalar boundary construction explains how averaging across moving routes can create a cusp; the CNN experiment tests finite route mediation in its own setting.

The 128-update branch forecast missed its calibration targets, and the rate selector missed its transfer criterion even though its mean held-out regret was lower. The next rate-decision study should test whether route-aware response improves signed loss, accuracy, or rate choice on newly locked

data and architectures. Matched-compute comparisons with signed hypergradients, direct branches, schedule forecasts, and rate-plus-probe controls would show the value of that added measurement.

Ethics statement

The work studies optimizer diagnostics on public image-classification benchmarks and synthetic two-moons data. The contribution concerns measurement of optimization dynamics. The authors report zero competing interests.

Reproducibility statement

The appendices give proofs, experimental settings, seed-level results, controls, and costs. They distinguish the stopped campaign from the later instrument check and the independent 20-trajectory study. The package includes that study's protocol and pre-outcome receipt, the complete trajectory roster, source code, hashes, and a verifier that recomputes the events and exact inference without training. Saved-result replays check the earlier image experiments. Figure sources and rendering checks are also included. Failed decision rules remain in the record.

AI use statement

Generative AI tools assisted with polishing the paper only. The authors take responsibility for the paper's text, claims, and code.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024. doi: 10.48550/arXiv.2412.20553. URL https://arxiv.org/abs/2412.20553.

Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum further constrains sharpness at the edge of stochastic stability. *arXiv preprint arXiv:2604.14108*, 2026. doi: 10.48550/arXiv.2604.14108. URL https://arxiv.org/abs/2604.14108.

Atilim Gunes Baydin, Robert Cornish, David Martínez, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BkrsAzWAb.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018. doi: 10.48550/arXiv.1812.01718. URL https://arxiv.org/abs/1812.01718.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2103.00065. URL https://arxiv.org/abs/2103.00065.

Ashok Cutkosky, Aaron Defazio, and Harsh Mehta. Mechanic: A learning rate tuner. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/955499a8e2860ed746717c1374224c43-Abstract-Conference.html.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2209.15594. URL https://arxiv.org/abs/2209.15594.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7449–7479, 2023. URL https://proceedings.mlr.press/v202/defazio23a.html.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173, 2017. URL https://proceedings.mlr.press/v70/franceschi17a.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, 2018. doi: 10.48550/arXiv.1806.04910. URL https://arxiv.org/abs/1806.04910.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep linear networks beyond the edge of stability. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2502.20531. URL https://arxiv.org/abs/2502.20531.

Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. *arXiv preprint arXiv:2311.02076*, 2023. doi: 10.48550/arXiv.2311.02076. URL https://arxiv.org/abs/2311.02076.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019. doi: 10.48550/arXiv.1806.01316. URL https://arxiv.org/abs/1806.01316.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Wonyeol Lee, Hangyeol Yu, and Hongseok Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/b096577e264d1ebd6b41041f392eec23-Abstract.html.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. doi: 10.48550/arXiv.2003.02218. URL https://arxiv.org/abs/2003.02218.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552, 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules. *arXiv preprint arXiv:2503.12811*, 2025. doi: 10.48550/arXiv.2503.12811. URL https://arxiv.org/abs/2503.12811.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, 2015. URL https://proceedings.mlr.press/v37/maclaurin15.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.csail.mit.edu/papers/v21/17-678.html.

Luke Metz, C. Daniel Freeman, Samuel S. Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2022. doi: 10.48550/arXiv.2111.05803. URL https://arxiv.org/abs/2111.05803.

Paul Micaelli and Amos J. Storkey. Gradient-based hyperparameter optimization over long horizons. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/596dedf4498e258e4bdc9fd70df9a859-Abstract.html.

Edward Milsom, Ben Anson, and Laurence Aitchison. Function-space learning rates. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44225–44251, 2025. URL https://proceedings.mlr.press/v267/milsom25a.html.

Ranjita Naik, Anh D. Nguyen, and Pankaj Kumar Singh. Predicting deep neural network training outcomes from early training telemetry. *arXiv preprint arXiv:2608.03709*, 2026. doi: 10.48550/arXiv.2608.03709. URL https://arxiv.org/abs/2608.03709.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL http://ufldl.stanford.edu/housenumbers/.

Haiquan Qiu, You Wu, Yingjie Tan, Yaqing Wang, and Quanming Yao. Spectral alignment as predictor of loss explosion in neural network training. *arXiv preprint arXiv:2510.04202*, 2025. doi: 10.48550/arXiv.2510.04202. URL https://arxiv.org/abs/2510.04202.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011. URL https://jmlr.org/papers/v12/reid11a.html.

Vincent Roulet, Atish Agarwala, Jean-Bastien Grill, Grzegorz Swirszcz, Mathieu Blondel, and Fabian Pedregosa. Stepping on the edge: Curvature aware learning rate tuners. *arXiv preprint arXiv:2407.06183*, 2024. doi: 10.48550/arXiv.2407.06183. URL https://arxiv.org/abs/2407.06183.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision*, pp. 464–472, 2017. doi: 10.1109/WACV.2017.58. URL https://arxiv.org/abs/1506.01186.

Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. doi: 10.48550/arXiv.1803.09820. URL https://arxiv.org/abs/1803.09820.

Lei Wu and Weijie J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2305.17490. URL https://arxiv.org/abs/2305.17490.

Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger Grosse. Understanding short-horizon bias in stochastic meta-optimization. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1803.02021. URL https://iclr.cc/virtual/2018/poster/240.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747. URL https://arxiv.org/abs/1708.07747.

A TANGENT REPRESENTATION AND TERMINAL SCALAR BENCHMARK

The forced recursion is standard forward hyperparameter differentiation (Franceschi et al., 2017). The following statements record the exact assumptions used here. Products with zero factors are interpreted as the identity.

Theorem 2 (Finite-horizon response representation and bound)**.** *Fix $T \geq 1$. Consider $\theta_{k+1} = G_{\eta,k}(\theta_k)$, where each map $(\theta, \eta) \mapsto G_{\eta,k}(\theta)$ is continuously differentiable and θ_0 is independent of η. Let $s_k = \partial_\eta \theta_k$ and*

$$A_k = D_\theta G_{\eta,k}(\theta_k), \qquad b_k = \partial_\eta G_{\eta,k}(\theta_k).$$

Then $s_0 = 0$, $s_{k+1} = A_k s_k + b_k$, and

$$s_T = \sum_{j=0}^{T-1} A_{T-1} \cdots A_{j+1} b_j. \tag{18}$$

Suppose $B \geq 0$, $C \geq 1$, $0 < \rho < 1$, and $\|b_j\| \leq B$. Also suppose $\|A_{k-1} \cdots A_j\| \leq C\rho^{k-j}$ for $0 \leq j < k \leq T$. Then

$$\|s_T\| \leq \frac{CB(1-\rho^T)}{1-\rho}. \tag{19}$$

Equation (18) decomposes the response into forced tangent products. For fixed B, C, T, its upper bound tends to CBT as $\rho \uparrow 1$. The realized response also depends on cancellation among forcing terms and their projection into predictor space.

Proposition 6 (Terminal scalar sensitivity–loss equivalence)**.** *Let $\theta_{t+1} = (1-\eta\lambda)\theta_t$, where $\lambda > 0$, $0 < |1-\eta\lambda| < 1$, and $\theta_0 \neq 0$. For an integer $T \geq 2$, define $L_T(\eta) = \frac{1}{2}\theta_T^2$ and $S_T(\eta) = |\partial_\eta \theta_T|$. Then*

$$S_T(\eta) = T\lambda|\theta_0|^{1/T}\{2L_T(\eta)\}^{(T-1)/(2T)}. \tag{20}$$

On every candidate set in the stable region, S_T and L_T induce the same preorder.

The equivalence concerns terminal Euclidean sensitivity. Since larger S_T accompanies larger L_T, this scalar example ranks high-sensitivity rates as high-loss rates. Fisher weighting and checkpoint maxima define different statistics.

B PROOFS

Proof of Theorem 2. Since θ_0 is independent of η, we have $s_0 = 0$. The chain rule gives

$$s_{k+1} = A_k s_k + b_k.$$

Repeated substitution yields

$$s_T = b_{T-1} + A_{T-1}b_{T-2} + \cdots + A_{T-1} \cdots A_1 b_0,$$

which is Equation (18) under the stated empty-product convention.

For the norm bound, index each product by its length $\ell = T - j - 1$. The empty product has norm one, which is bounded by C. The remaining products satisfy the contraction hypothesis. Hence

$$\begin{aligned}\|s_T\| &\leq \sum_{j=0}^{T-1} \|A_{T-1} \cdots A_{j+1}\| \, \|b_j\| \\ &\leq CB \sum_{\ell=0}^{T-1} \rho^\ell = \frac{CB(1-\rho^T)}{1-\rho}.\end{aligned}$$

This proves both claims. □

Proof of Proposition 6. Set $\mu = 1 - \eta\lambda$. The recursion gives

$$\theta_T = \mu^T \theta_0, \qquad \partial_\eta \theta_T = -T\lambda\mu^{T-1}\theta_0.$$

Therefore

$$S_T = T\lambda|\theta_0||\mu|^{T-1}, \qquad 2L_T = |\theta_0|^2|\mu|^{2T}.$$

The second identity implies

$$|\mu|^{T-1} = \left(\frac{2L_T}{|\theta_0|^2}\right)^{(T-1)/(2T)}.$$

Substitution gives Equation (20). Its prefactor is positive, and its exponent on L_T is positive. Thus the two quantities induce the same preorder. □

Proof of Proposition 1. Fix k and write $\theta = \theta_k(\eta)$. For a parameter increment h, define

$$\Phi_{\rightarrow}(h) = \mathbb{E}_Q D_{\mathrm{KL}}(p_\theta(\cdot \mid x) \| p_{\theta+h}(\cdot \mid x)).$$
$$\Phi_{\leftarrow}(h) = \mathbb{E}_Q D_{\mathrm{KL}}(p_{\theta+h}(\cdot \mid x) \| p_\theta(\cdot \mid x)).$$

Regularity permits differentiation under the expectation. The score and information identities give

$$\Phi_{\rightarrow}(0) = \Phi_{\leftarrow}(0) = 0, \qquad \nabla\Phi_{\rightarrow}(0) = \nabla\Phi_{\leftarrow}(0) = 0,$$

and

$$\nabla^2\Phi_{\rightarrow}(0) = \nabla^2\Phi_{\leftarrow}(0) = F_Q(\theta).$$

Taylor expansion at the origin therefore yields

$$\Phi_{\rightarrow}(h) = \frac{1}{2}h^\top F_Q(\theta)h + o(\|h\|^2),$$
$$\Phi_{\leftarrow}(h) = \frac{1}{2}h^\top F_Q(\theta)h + o(\|h\|^2).$$

The average of the two directional expansions gives the same formula for D_{sym}. Now set $h = \theta_k(\eta+\delta) - \theta_k(\eta) = \delta s_k + o(\delta)$. Then $\|h\| = O(|\delta|)$ and

$$h^\top F_Q(\theta)h = \delta^2 s_k^\top F_Q(\theta)s_k + o(\delta^2).$$

Substitution and division by δ^2 prove the result. The same argument applies pointwise on a fixed empirical probe and then averages over its examples. □

B.1 Signed response and trajectory curvature

Proof of Proposition 4. Let $A(f) = \log\sum_c \exp(f_c)$. Softmax cross-entropy is $\ell_y(f) = A(f) - e_y^\top f$, while direct substitution into the categorical KL gives

$$D_{\mathrm{KL}}(p\|\widetilde{p}) = A(\widetilde{f}) - A(f) - p^\top(\widetilde{f} - f).$$

Adding $(p - e_y)^\top(\widetilde{f} - f)$ proves Equation (12) exactly.

For the differential statements, $\nabla_f\ell_y = p - e_y$ and $\nabla_f^2\ell_y = \mathrm{diag}(p) - pp^\top$. The first-order chain rule gives Equation (13). Applying the second-order chain rule to each example gives

$$\nabla_a^2\ell_{y_i}(f_i(a_0)) = J_i^\top F_i J_i + \sum_c (p_{ic} - e_{y_i,c})H_{ic}.$$

Averaging proves Equation (14). □

Stable two-step SGD example. Train a scalar w on $w^2/2$ from $w_0 = 1$ for two gradient steps with $\eta = e^u$. Then $w_2 = (1-\eta)^2$. Use validation logits $(0, w_2)$ and label one, so $p = \sigma(w_2)$. At $\eta = 3/4$, which lies inside the scalar stability interval $(0, 2)$,

$$w_2 = \tfrac{1}{16}, \qquad \partial_u w_2 = -\tfrac{3}{8}, \qquad \partial_u^2 w_2 = \tfrac{3}{4}.$$

Consequently,

$$G_V = p(1-p)(\partial_u w_2)^2 = 0.03512194, \qquad C_V = (p-1)\partial_u^2 w_2 = -0.36328506,$$

and $\partial_u^2 L_V = G_V + C_V = -0.32816312$. Trajectory curvature therefore reverses the sign suggested by the positive Fisher term in this scalar calculation.

Coordinate qualification. For a smooth coordinate change $a = g(b)$, the ordinary Hessian satisfies

$$\nabla_b^2 L_V = Dg^\top (\nabla_a^2 L_V) Dg + \sum_c (\partial_{a_c} L_V) \nabla_b^2 g_c.$$

The Fisher term transforms as $G_V^b = Dg^\top G_V^a Dg$, so C_V acquires the same gradient-dependent correction as the Hessian. The displayed negative scalar curvature is in the declared $u = \log \eta$ coordinate.

B.2 Proof of the two-update boundary mechanism

Proof of Proposition 5. The training loss is $\log(1+\exp(\mathrm{ReLU}(w)))$. Away from zero its SGD map is

$$\Phi_\eta(v) = \begin{cases} v - \eta\sigma(v), & v > 0, \\ v, & v < 0. \end{cases}$$

The convention at zero fixes the map there; the equality event has probability zero under f. Since $W > 0$, put $u = w - \eta\sigma(w)$. Then

$$w_2(w, \eta) = \begin{cases} u - \eta\sigma(u), & \eta < T(w), \\ u, & \eta > T(w). \end{cases}$$

On $w \in (1, 6/5)$ and $\eta \in (7/5, 3/2)$, $u > 0$ implies $u < 6/5 - (7/5)/2 = 1/2$, hence $w_2 < 1/2 - 7/10 = -1/5$. If $u < 0$, $w_2 = u < 0$. The probe logit is consequently $z = -w_2 > 0$ on both open branches, with no further probe activation surface. Both formulas extend smoothly to their one-sided limits:

$$\lim_{\eta \uparrow T(w)} w_2 = -T(w)/2, \qquad \lim_{\eta \downarrow T(w)} w_2 = 0.$$

Thus the probe probabilities approach $\sigma(T(w)/2)$ and $1/2$. At $x = 1$, both second-step states instead give probe probability $1/2$; the negative-input probe is essential.

The threshold map satisfies $T(w_*) = \eta_*$ and $T'(w) = 1 + (1 - w)e^{-w} > 0$ on the support rectangle. Its derivative is bounded away from zero on the compact support. Change of variables gives the stated threshold density, which is C^1 near η_*. For $h > 0$, precisely $T(W) \in (\eta_*, \eta_* + h)$ crosses the boundary; for $h < 0$, the interval is $(\eta_* + h, \eta_*)$. Its probability is $\rho_T(\eta_*)|h| + O(h^2)$.

For Bernoulli logits a, b the declared divergence is

$$D_{\mathrm{sym}}(\mathrm{Ber}(\sigma(a)), \mathrm{Ber}(\sigma(b))) = \tfrac{1}{2}[\sigma(a) - \sigma(b)](a - b).$$

On the crossing set, $w - w_* = O(|h|)$ and the two logits tend uniformly to $\eta_*/2$ and zero. The divergence is therefore $J(\eta_*) + O(|h|)$. Multiplying by the crossing probability gives $\rho_T(\eta_*)J(\eta_*)|h| + O(h^2)$, with a positive coefficient. Symmetry gives the same leading coefficient for either sign of h, without implying exact evenness.

On a noncrossing initialization, the complete rate segment lies in one smooth branch. With $s = \sigma(w)$, its logit derivative is

$$z_\eta = \begin{cases} s[1 - \eta\sigma'(u)] + \sigma(u), & u > 0, \\ s, & u < 0. \end{cases}$$

Its magnitude is at most two. Logistic is $1/4$-Lipschitz, so $D_{\mathrm{sym}} \le |\Delta z|^2/8 \le h^2/2$. The integrated noncrossing response is $O(h^2)$, proving Equation (15). For any fixed $w \neq w_*$, sufficiently small h remains in one branch and the ordinary expansion yields

$$q(w) = \tfrac{1}{2}\sigma(z)[1 - \sigma(z)]z_\eta^2 \le \tfrac{1}{2}.$$

The quotients on the crossing set are of order h^{-2} on a set of mass $O(|h|)$; bounded pathwise derivatives therefore do not provide an integrable domination of the finite quotients. This proves the noninterchange claim. □

Local Jacobians and global discontinuity. Each smooth SGD branch has derivative $1 - \eta\sigma'(v)$ for $v > 0$ and one for $v < 0$, both in $[5/8, 1]$. The cusp arises while these local products remain bounded. The global update has limits zero and $-\eta/2$ at zero, making the training-gradient update discontinuous. The negative-input probe detects the resulting state jump.

Expected labeled-loss consequence. For a fixed probe label $y \in \{0,1\}$ define $L_y(\eta) = \mathbb{E}_W[\log(1+e^{z(W,\eta)}) - yz(W,\eta)]$. Changing variables to $t = T(W)$ separates the two smooth integrals at $t = \eta$. The Leibniz rule, for η near η_*, gives

$$L_y'(\eta) = \mathbb{E}_W[(\sigma(z) - y)z_\eta] + \rho_T(\eta)\left\{\log 2 - \log(1+e^{\eta/2}) + y\eta/2\right\}. \tag{21}$$

The second term is the loss after the jump minus the loss before it, weighted by threshold density. Boundary contributions to derivatives of expectations are established in nonsmooth reparameterization (Lee et al., 2018); this is an explicit learning-rate specialization. It concerns the averaged objective and leaves conditional differentiation of a fixed smooth trajectory intact.

For W uniform on $[\log 3 - .05, \log 3 + .05]$, the local density satisfies the proposition's assumptions. At η_*, the mean pathwise and full derivatives are $+.518633$ and -3.946227 for $y = 0$, and $-.345440$ and $+2.762713$ for $y = 1$. Split deterministic quadrature and central differences agree within 6×10^{-10} for these expected derivatives. The cusp coefficient is .6639014472, while $\mathbb{E}q(W) = .0878787808$. An independent check with a normalized C^1 compact bump density also recovers both claims.

Finite samples and intervention laws. For a fixed finite sample $W_1, \ldots, W_n$, almost surely $\min_i |T(W_i) - \eta_*| > 0$. Below that distance, every sampled trajectory stays in its smooth branch and the empirical response divided by h^2 converges to the empirical mean of $q(W_i)$. Estimating the population cusp requires initialization replications that resolve threshold crossings at the chosen increment. For bounded Z independent of W, uniformity of the local remainder gives $\mathbb{E}_Z R(aZ) = \rho_T(\eta_*)J(\eta_*)a\mathbb{E}|Z| + O(a^2\mathbb{E}Z^2)$. The leading coefficient under $\eta = g(b)$ acquires $|g'(b_*)|$, consistent with its linear order; the within-branch Fisher coefficient retains its squared-Jacobian transformation.

The half-sum symmetric KL and shared initialization are part of the statement. Forward KL generally has different one-sided leading coefficients. Fixed smooth softplus removes this particular update discontinuity. The argument takes the increment limit at fixed ReLU nonlinearity.

Proof of Proposition 2. The chain rule gives

$$s_k^u = \partial_u \theta_k = \frac{d\eta}{du} s_k^\eta.$$

Substitution into Equation (3) gives Equation (7). At a fixed candidate rate, the Jacobian factor is independent of k. It therefore also factors out of the maximum-over-time coefficient.

For the ranking claim, fix two rates $\eta_1 < \eta_2$ with coefficients $a_1, a_2 > 0$. Write $q_i = h'(\eta_i)$. Their transformed coefficients are a_i/q_i^2. Positive choices of q_1, q_2 can produce either strict ordering. A smooth positive function q can interpolate these two values. Defining

$$h(\eta) = h(\eta_0) + \int_{\eta_0}^{\eta} q(t)\, dt$$

produces an increasing smooth diffeomorphism with the chosen derivatives. This proves the ranking claim. □

Proof of Proposition 3. Equation (7) and $du = h'(\eta)d\eta$ give

$$\sqrt{\mathcal{S}_T^{F,u}(h(\eta))}\, du = \sqrt{\left(\frac{1}{h'(\eta)}\right)^2 \mathcal{S}_T^{F,\eta}(\eta)}\, h'(\eta)d\eta = \sqrt{\mathcal{S}_T^{F,\eta}(\eta)}\, d\eta.$$

Integration over the two representations of the same physical interval proves the identity. □

C Finite-intervention certificates and checkpoint quantifiers

Proof of Theorem 1. At a base tie the stated bounds are trivial. Otherwise, if the selected class under q_x is $j \neq c$, then $q_j \geq q_c$ and

$$\kappa \leq p_c - p_j = (p_c - q_c) + (q_c - q_j) + (q_j - p_j) \leq |p_c - q_c| + |p_j - q_j| \leq \|p - q\|_1.$$

Pinsker's inequality with natural logarithms gives both $D_{\mathrm{KL}}(p\|q) \geq \|p-q\|_1^2/2$ and $D_{\mathrm{KL}}(q\|p) \geq \|p-q\|_1^2/2$, so their average obeys the same bound (Reid & Williamson, 2011). Thus a prediction change implies $D_\diamond \geq \kappa^2/2$. The pointwise indicator is bounded by $\min\{1, 2D_\diamond/\kappa^2\}$. Integrating gives the second statement. Splitting the disagreement event according to $\kappa \leq \gamma$ gives Equation (16). Infinite divergences simply make a bound trivial. For any joint distribution of (X, Y) with marginal Q, the two zero-one losses agree whenever the predicted classes agree; their expected absolute difference is therefore bounded by the same disagreement probability. □

Kernel and coordinate invariance. For measurable predictors and a fixed kernel K_η, apply the theorem conditionally on (ω, η') and average. The margin tail belongs to the base predictor and does not depend on the alternative rate. Under a measurable bijection h with measurable inverse, drawing $u' \sim h_\# K_\eta$ is the same as drawing $\eta' \sim K_\eta$ and setting $u' = h(\eta')$. Both physical predictors, divergences, margins, and disagreement indicators are unchanged. Their integrals are therefore unchanged. This proves the invariance claim in Section 4 without a differentiability assumption on h.

Checkpoint aggregation. For a fixed physical pair, write F_k for disagreement, $R_k = \mathbb{E}_Q D_k$, and $M_k(\gamma) = Q\{\kappa_k \leq \gamma\}$. For fixed weights $w_k \geq 0$ summing to one,

$$\sum_k w_k Q(F_k) \leq \min\{1, \sum_k w_k M_k(\gamma) + 2\sum_k w_k R_k/\gamma^2\}.$$

Also $\max_k Q(F_k) \leq \min\{1, \max_k M_k(\gamma) + 2R_{\max}/\gamma^2\}$. These statements concern a drawn checkpoint or the worst checkpoint probability. In contrast, an input changing prediction at *any* of T checkpoints obeys

$$\begin{aligned} Q\Big(\bigcup_{k=1}^{T} F_k\Big) &\leq \min\{1, Q(\min_k \kappa_k \leq \gamma) + 2\mathbb{E}_Q \max_k D_k/\gamma^2\} \\ &\leq \min\{1, Q(\min_k \kappa_k \leq \gamma) + 2TR_{\max}/\gamma^2\}. \end{aligned} \tag{22}$$

Indeed, outside the low-margin event, any change forces $\max_k D_k \geq \gamma^2/2$; then use $\mathbb{E}\max_k D_k \leq \sum_k \mathbb{E}D_k \leq TR_{\max}$. Dropping the factor T by interchanging expectation and maximum is unjustified.

D Finite-quotient error and sufficient derivative bounds

Error and rank proof. Under the hypotheses preceding Equation (17), Taylor's integral remainder gives $|\phi_{ik}(\delta_i) - q_{ik}\delta_i^2| \leq M_{ik}|\delta_i|^3/6$. Division proves the displayed bound. Since maximum is 1-Lipschitz in the supremum norm,

$$|\max_k \phi_{ik}(\delta_i)/\delta_i^2 - \max_k q_{ik}| \leq \max_k e_{ik}.$$

The triangle inequality proves the weighted-average version and the strict-gap criterion. Any independently certified numerical error must be added to each e_{ik}. A new coordinate requires its own error calculation: for $a_\delta = \delta/[h(\eta+\delta) - h(\eta)]$ and $a_0 = 1/h'(\eta)$, the exact finite quotient has error at most $a_\delta^2 e + |a_\delta^2 - a_0^2|q$ relative to the transformed local coefficient.

Explicit sufficient conditions. Fix one candidate, checkpoint, and admissible interval $[\eta - h, \eta + h]$, $h > 0$. Suppose the finite-class probabilities are positive and measurable in x, and $\ell_j(r, x) = \log p_r(j \mid x)$ is C^3 in r with uniform bounds

$$|\ell_j'| \leq A, \qquad |\ell_j''| \leq B, \qquad |\ell_j'''| \leq C$$

for Q-almost every input, all classes, and the whole interval, where A, B, C are finite. For $0 < |\delta| \leq h$, valid constants in Equation (17) are

$$\begin{aligned} &M_\rightarrow = C, \\ &M_\leftarrow = hA(A^3 + 3AB + C) + 2A^3 + 3AB, \qquad M_{\mathrm{sym}} = \tfrac{1}{2}(M_\rightarrow + M_\leftarrow). \end{aligned} \tag{23}$$

To verify this, evaluate $a_j = \ell_j'$, $b_j = \ell_j''$, and $c_j = \ell_j'''$ at $\eta + r$, and set $d_j = \ell_j(\eta + r) - \ell_j(\eta)$. Then $|d_j| \le hA$ and differentiation gives

$$\phi_{\rightarrow}'''(r) = -\mathbb{E}_Q \sum_j p_j(\eta) c_j, \quad \phi_{\leftarrow}'''(r) = \mathbb{E}_Q \sum_j p_j(\eta + r)[(a_j^3 + 3a_j b_j + c_j)d_j + 2a_j^3 + 3a_j b_j].$$

The bounded derivatives justify interchange with the input expectation. Normalization $\sum_j p_j = 1$ shows that both divergences and first derivatives vanish at zero and both second derivatives equal $\mathbb{E}_Q \sum_j p_j(\eta)\{\ell_j'(\eta)\}^2 = 2q$. Equation (23) and Taylor's theorem follow. In particular, averaging the two directional divergences at the same endpoints does not eliminate the cubic remainder generally.

Dependence on trajectory and horizon. Suppose along every trajectory in the interval that $\|\theta'\| \le V_1$, $\|\theta''\| \le V_2$, $\|\theta'''\| \le V_3$, and the first three parameter derivatives of $\log p_\theta(j \mid x)$ have operator norms bounded by G_1, G_2, G_3. Then valid choices are

$$A = G_1 V_1, \quad B = G_2 V_1^2 + G_1 V_2, \quad C = G_3 V_1^3 + 3G_2 V_1 V_2 + G_1 V_3.$$

For a C^3 augmented-state map $G_k(z, r)$ with uniform derivative bounds $\|D_z G_k\| \le a_k$, $\|\partial_r G_k\| \le b_k$, $\|D^2_{(z,r)} G_k\| \le c_k$, and $\|D^3_{(z,r)} G_k\| \le d_k$ in the product norm $\|(v, t)\| = \|v\| + |t|$, use zero initial derivative bounds and propagate

$$\begin{aligned} V_{1,k+1} &= a_k V_{1,k} + b_k, \\ V_{2,k+1} &= a_k V_{2,k} + c_k (V_{1,k} + 1)^2, \\ V_{3,k+1} &= a_k V_{3,k} + 3c_k (V_{1,k} + 1) V_{2,k} + d_k (V_{1,k} + 1)^3. \end{aligned}$$

The chain rule applied to $(z_k(r), r)$ proves these bounds; the usual parameter projection has operator norm one. The constants make horizon dependence explicit. Applying them to an image run requires uniform derivative control along its rate interval, including any activation or guarded-update branches; that information is absent from the archived records.

E CONTROLLED EVIDENCE DETAILS

Normal-form mechanism check. The scalar normal-form control uses $U(x) = \frac{1}{2}x^2 - \frac{1}{4}x^4 + \frac{0.2}{6}x^6$, $x_0 = 0.15$, and 260 full-batch gradient steps. Its reported response is $\max_k s_k^2$, where $s_{k+1} = (1 - \eta U''(x_k))s_k - U'(x_k)$. This is the exact Euclidean tangent response. The fixed point has multiplier $1 - \eta$, and the response maximum lies next to its known flip on the tested grid. The example shows a large forced response near a known scalar transition.

MNIST MLP seed-level alignment. Table 3 exposes the four curves behind the aggregate in Table 4. In every seed, the stable-prefix, additive-KL-peak, and best-validation rates are the same grid point; the first divergent rate is the next grid point. The alignment holds at every tested seed on this grid.

Table 3: MNIST MLP seed-level sweep. The stable prefix, KL peak, and validation reference agree within every seed. Divergence begins at the next grid point.

Seed	First divergence	Stable prefix	KL peak	Best validation
0	.7848	.7067	.7067	.7067
1	.8629	.7848	.7848	.7848
2	.9410	.8629	.8629	.8629
3	.8629	.7848	.7848	.7848

Figure 2 shows the aggregate full-batch two-moons sweep and the separation between its susceptibility, validation, and curvature references.

MNIST MLP shadow trajectories. Synchronized shadows under mini-batch SGD give the same within-seed alignment: the additive KL peak, stable-prefix endpoint, and best validation rate coincide, with divergence beginning at the next tested rate (Tables 4 and 3).

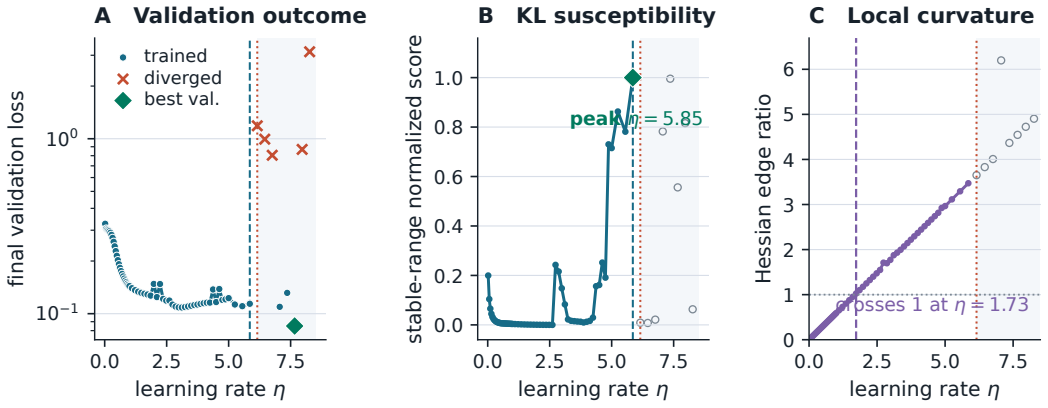


Figure 2: Controlled full-batch two-moons MLP sweep. The additive Fisher/KL sensitivity peak occurs at the stable-prefix endpoint (blue dashed line); local curvature crosses one earlier. The orange dotted line and gray band begin at the first divergence, while open points show recorded post-prefix diagnostics. The green diamond marks the best final validation loss, which is outside the stable prefix. Final loss uses a log axis; panel B uses range normalization over the prefix.

MNIST CNN separation. The compact CNN separates all four grid-defined references. Curvature appears first. The prediction-space response follows, then the validation reference and stable-prefix endpoint. Table 4 gives their locations. Appendix F studies regimes with different alignments.

Table 4: Grid-defined sweep references. The stable prefix ends before the first divergence. Sensitivity is $\max_k s_k^2$ in the normal form and η_F in neural sweeps. Validation and curvature provide endpoint and local references.

Setting	Flip/stable prefix	Sensitivity peak (η)	Validation	Curvature
Normal form	2.00	2.02	NA	2.00
Small MLP	5.85	5.85	7.66	1.73
MNIST MLP	.785	.785	.785	NA
MNIST CNN	2.22	.220	.358	.068

F Archived selection, proxy, and width evidence

Table 5: Selector summaries. Each method cell gives the mean selected rate followed by mean $\log_2$ error in parentheses. "Target" is the mean final-validation reference rate; means need not be grid points.

Setting	Target	KL	Probe	Combined
Fashion-MNIST CNN	.378	.170 (1.140)	.290 (.469)	.306 (.533)
MNIST CNN	.408	.254 (.671)	.360 (.207)	.360 (.207)
MNIST CNN, momentum	.070	.060 (.678)	.127 (.548)	.127 (.548)
MNIST CNN, Adam	.013	.001 (3.636)	.013 (.000)	.008 (.667)
SVHN TinyResNet	.017	.005 (1.667)	.010 (.667)	.017 (.667)

Selector definition and scope. The archived rule minimizes $J(\eta) = \widetilde{L}_{\mathrm{probe}}(\eta)^2 + 0.1\widetilde{R}_{\mathrm{KL}}(\eta)^2$, where both terms are normalized over the candidate grid and $R_{\mathrm{KL}} = \log(1 + \widehat{S}_{\max}^{\mathrm{KL},\eta})$. Table 5 compares selected rates with the retrospective final-validation grid optimum; error is mean seed-level absolute $\log_2$ distance. Probe loss has the lowest mean error in all five image/optimizer settings. Adding KL ties it in three settings and worsens it in two. Equal mean selected rates can conceal seed-level differences: for MNIST Adam, probe loss has zero mean error, while the combined rule has .667.

Eligibility in the archived producer uses a final-status stable prefix. All 11 failures in the saved breadth collection occur at steps 5–38, before the 40-step probe. Thus their failure status was available by the probe; future use of the rule should determine eligibility at that same decision point.

Homogeneous-proxy check. The synthetic E1 benchmark uses a homogeneous tangent-growth proxy Λ_T, omitting the forced response and Fisher geometry of the main instrument. Across 32 seed/problem clusters, there are 168 probe survivors and 16 later unstable runs. The proxy AUC is .6978, compared with .7385 for probe loss. The seed-cluster bootstrap difference is $-.0407$, with 95% interval $[-.2032, .0962]$; the proxy leads in 27.0% of resamples. Probe loss has the stronger observed AUC in this synthetic proxy comparison.

Mini-batch control. Across five two-moons seeds the stable-prefix rate is $2.17 \pm .23$ and the additive KL maximum is $.84 \pm 1.03$. Stochastic dynamics can separate these references substantially.

F.1 E3 AND R3 ARE SEPARATE EXPLORATORY SETTINGS

The six-rate E3 grid records no divergence in 24 width-three runs, leaving its transition unbracketed. R3 observes ordered references at width 48 in four seeds. The network, optimizer, data subsets, and diagnostic also change (Table 6), so R3 is a separate exploratory width setting.

Table 6: Configuration differences between archival width studies. Both validation sources are the official SVHN test split. R3 changes the grid alongside the listed design choices.

Item	E3	R3
Network	Five residual blocks; channels 48, 96, 192	Two residual blocks; channels 48, 96 (also width 16)
Normalization	GroupNorm	BatchNorm
Downsampling	Two stages	One stage
Optimizer	Momentum .9; weight decay .0005	Plain SGD
Data subsets	Random 4096/1024	First 8192/4096
Probe / total steps	80 / 500	60 / 400
Rate grid	Six rates, .008–.048	Fifteen rates, .01–8.0
Curvature statistic	One-step Rayleigh quotient	Twelve-step Hessian-vector norm

At width 48, the (curvature, additive KL maximum, first recorded divergence) triples are $(.02, 1.6, 2.6)$, $(.01, .2, 3.5)$, $(.02, 1.2, 2.6)$, and $(.02, .05, 3.5)$. They are strictly ordered in this setting. The runner's general ordering predicate is non-strict; at width 16, two of four triples tie the curvature and KL locations. The KL maximum is a curve statistic; the first recorded divergence is a separate runner diagnostic. The ordering describes R3 under its own configuration.

R3's Hessian-vector norm estimates a largest-magnitude mode when power iteration converges, including possible negative curvature. Its failure flag combines nonfinite base or shadow minibatch loss with a base validation-loss cutoff of 20 checked every 100 updates. The saved scalar rows locate that compound runner flag; they contain no final 400-update check or base-only numerical-divergence trace.

G DATA LINEAGE AND EXPERIMENTAL PROTOCOL

G.1 SVHN STAGES, MATCHING, AND ARCHIVED CRITERIA

For display only, define the monotone transforms

$$A_\eta = \log_{10}(1 + \widehat{\mathcal{S}}^{\mathrm{KL},\eta}_{\max}), \qquad A_{\log \eta} = \log_{10}(1 + \eta^2 \widehat{\mathcal{S}}^{\mathrm{KL},\eta}_{\max}). \tag{24}$$

An eight-seed development sweep uses rates $\{.008, .012, .017, .024, .034, .048\}$, an 80-step probe, and 500 updates. Mean probe loss remains within a 0.03 matching band, while later loss rises at the high end. All seeds have opposing coordinate contrasts between the endpoint rates; individual full curves need not be monotone. Within that band, the minimum and maximum of $z(\mathrm{SD}_{\mathrm{val}}) + z(\mathrm{worst}_{\mathrm{val}}) + 0.5z(\mathrm{mean}_{\mathrm{val}})$ select the pair $.008, .048$.

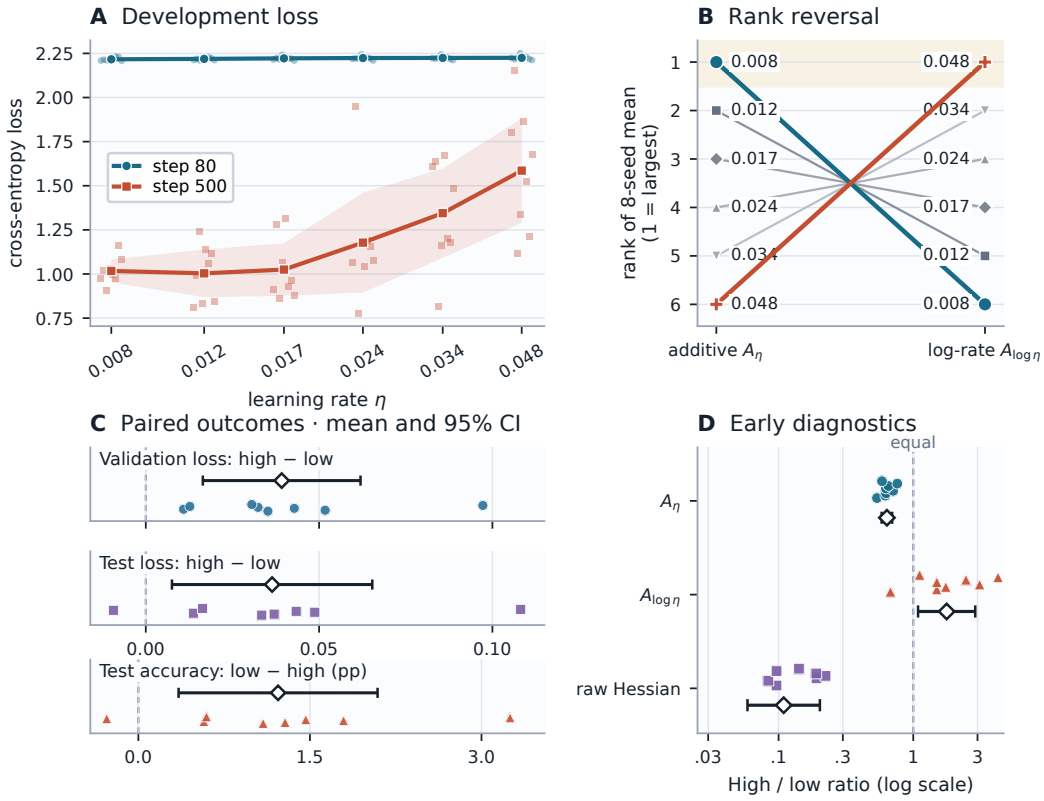


Figure 3: SVHN coordinate audit ($n = 8$ matched seeds). (A) Raw development seed points, means, and two-sided 95% t intervals. (B) Rank changes for the six mean scores in Equation (24); rank 1 is largest. (C) All paired full-data outcomes on separate loss and accuracy scales; diamonds and bars show paired means and 95% t intervals, and positive values favor $\eta = .008$. (D) Per-seed high/low diagnostic ratios on a log scale; diamonds and bars show geometric means and 95% t intervals computed in log-ratio space. The development procedure uses official-test validation, so later official-test outcomes are descriptive.

Table 7: Full-data SVHN paired effects ($n = 8$). Intervals and two-sided p-values use paired t calculations, conditional on the fixed data and selected pair; they are unadjusted for multiple endpoints or prior selection. Scores use Equation (24).

Contrast	Mean	95% interval	Positive pairs	p
Probe loss: high − low	.00919	[.00445, .01392]	7/8	.0025
Final validation loss: high − low	.03921	[.01647, .06195]	8/8	.0047
Official test loss: high − low	.03645	[.00755, .06534]	7/8	.0204
Official test accuracy: low − high	.01222	[.00352, .02091]	7/8	.0128
Additive score: low − high	1.27042	[1.02525, 1.51558]	8/8	$< 10^{-4}$
Log-rate score: high − low	.06732	[.00189, .13274]	7/8	.0452
Raw Hessian: low − high	6.14912	[4.09509, 8.20315]	8/8	.0002

Development uses eight seeds 0–7, six rates, 4096 training and 1024 validation examples, split seed 123, batch size 128, momentum .9, weight decay .0005, an 80-step probe and 500-step endpoint. KL uses 768 validation examples and symmetric divergence. The source loader constructs validation from the official test split. The saved CSV lacks an executed index manifest; the documented producer path supplies the evidence for this data reuse. Under those settings, the 1024 development validation examples are among the later full test set.

The full-data pair $.008, .048$ uses eight seeds 0–7, the official training split with a 5000-example validation holdout (split seed 2026), batch size 128, momentum .9, weight decay $.0005$, 8000 updates, and an 80-step probe. KL uses 768 validation examples; Hessian telemetry uses 96 examples and two power iterations. The test split is evaluated at the full-data endpoint after serving as a development validation source under the documented procedure. The size and direction of selection bias remain unquantified.

An archived local protocol dated June 6, 2026 fixes the pair, seeds, horizon and 5% perturbation; it has no independent timestamp. It declares mean probe-loss matching within .15 and later higher mean *or* variance at the high rate, with test direction and diagnostic order examined. These written directional criteria differ from the source's minimum-effect Boolean diagnostics. Table 8 preserves both definitions.

Table 8: Written directional criteria alongside runner diagnostics with additional minimum-effect thresholds.

Quantity	Observation / requirement	Status
Mean probe loss gap	.00919 / at most .15	Written tolerance satisfied
Final validation mean, low / high	.21914 / .25834; high larger	Written direction satisfied
Final validation SD, low / high	.00974 / .02273	Direction agrees
Maximum validation-loss gap or SD gap	$.06715 < .15$; $.01299 < .075$	Runner flag false
Mean log-rate display-score gap	$.06732 < .10$	Runner flag false

The $.00919$ probe gap meets the declared tolerance and also differs from exact equality under the paired t test ($p = .0025$). The paired intervals describe seed variation for a selected pair on fixed data and are unadjusted for selection or multiple endpoints. For the log-rate contrast, the exact two-sided sign-test p for seven of eight positive pairs is $.0703$, compared with paired t $p = .0452$. The two tests describe different aspects of this eight-seed contrast.

Shadow validity and curvature. Archived image producers guard the shadow branch and may skip an update after a nonfinite or large shadow minibatch loss. Their output tables retain neither skip counts nor per-checkpoint KL and maximum times. We therefore analyze their saved scores as finite responses under the recorded producer, with coordinate renormalization applied directly to those scores. The new smooth calibration records full status traces and runs ordinary SGD throughout. SVHN development and paired Hessian telemetry use one and two power iterations, respectively.

Other archived settings. MNIST/Fashion-MNIST functional-selector CNNs under SGD use five seeds, batch size 256, a 40-step probe, and 120-step endpoint. The MNIST momentum and Adam variants each use three seeds. SVHN TinyResNet selectors use three seeds, batch size 128, momentum .9, weight decay $.0005$, a 120-step probe, and 400-step endpoint. Their standard subset size is 4096/1024. MLP shadow sweeps use four seeds, batch size 256, and 130 updates. Every retained coordinate-audit row has relative increment .05; the minimum-step floor is inactive in the retained rows. Stable-prefix response peaks are retrospective grid statistics. The package preserves the archived raw tables and runners.

G.2 Frozen prospective selector and held-out decision

The v2 candidate grid is $\{.02, .05, .08, .12, .18, .27, .4, .6, .9, 1.3, 1.8, 2.4, 3.2\}$. Each small CNN uses batch size 128, ordinary SGD without momentum or weight decay, a 40-update probe, a 120-update endpoint, and a synchronized shadow at 1.05η. Within a dataset/seed block, all rates share initialization and minibatches. The scheduler carries partial permutation tails instead of discarding them. State is recorded from step zero. A nonfinite loss, gradient, or parameter permanently stops the affected trajectory; there is no threshold guard, skipped update, clipping, or rescue. Candidate eligibility is fixed at the probe and never reads endpoint status.

The official training split is partitioned with fixed dataset-specific seeds into 4096 training, 512 probe, and 1024 development-endpoint examples. Development uses MNIST and Fashion-MNIST seeds 0–4. Continuous features and the endpoint target are standardized within each candidate block. Leave-one-block-out development regret selects ridge penalties 100 for rate plus probe and .1 for

the full model. The serialized model SHA-256 begins `253ef81e245df5ef`. Held-out MNIST seeds 100–107, Fashion-MNIST seeds 200–207, and Kuzushiji-MNIST seeds 300–307 are disjoint from development; their endpoint is the complete official test split. The model and policy are locked before these held-out evaluations.

Table 9 reports the locked policy regrets and predeclared contrast on the held-out blocks.

Table 9: Frozen prospective endpoint regret. Δ is full minus rate-plus-probe regret; negative favors response and margin features. Each dataset has eight held-out model-seed blocks.

Dataset	Probe	Rate + probe	Full	Δ
MNIST	.0515	.0515	.0294	$-.0221$
Fashion-MNIST	.3333	.3445	.1575	$-.1870$
Kuzushiji-MNIST	.0781	.1020	.1291	$+.0270$
All 24 blocks	.1543	.1660	.1053	$-.0607$

The stratified block-bootstrap interval for the final row is $[-.1972, .0266]$. The full and rate-plus-probe models choose the same rate in 19 blocks; among the other five, the full model wins three and loses two. Mean absolute $\log_2$ rate error is .5287 for probe loss, .6255 for rate plus probe, and .6240 for the full model. The predeclared gate requires an upper interval endpoint below zero, improvement on at least two datasets, and no dataset degradation above .01. Improvement on two datasets is the sole criterion met. The 312 held-out candidate pairs, full traces, selections, bootstrap result, and recorded stop decision are in `code/prospective_selector_v2/results/`.

G.3 Post-hoc prospective diagnostics and development ablations

The following exploratory analyses use the saved prospective cohort after its decision was recorded. Candidate rates within a curve are dependent, so bootstrap resampling uses seed blocks within the fixed dataset strata.

Table 10 reports the post-hoc coordinate audit separately for the development and held-out cohorts.

Table 10: Separate prospective-cohort coordinate audits. The intervals condition on the fixed datasets and recipe. The development and held-out audits are both post hoc.

Cohort	Curves	Changed peak	Reversed pairs	Fraction (95% interval)
Development	10	7	185/780	23.72% [18.85, 28.85]
Held-out	24	17	406/1872	21.69% [20.25, 23.18]

The 13 rates are common to every curve, and every base/shadow pair completes. The relative coefficient is the additive finite coefficient multiplied by η^2; it therefore ranks the fixed-relative raw divergences identically. These are algebraic comparisons of saved predictor pairs. Selecting the additive or relative maximum on those records gives held-out mean endpoint regret .3603 or .9261, respectively, versus .1543 for the frozen probe selector. This post-hoc illustration uses the 24-curve prospective cohort, kept separate from the archival 19-curve collection.

Influence and informativeness. Full minus rate-plus-probe regret has three wins, 19 ties, and two losses. Fashion-MNIST seed 205 contributes -1.495881 to the sum of paired contrasts. Removing this single block as an influence diagnostic gives mean $+.001724$ over the remaining 23; the original analysis retains all 24. Against plain probe loss, the full model has two wins, 19 ties, and three losses, with mean contrast $-.048975$ and a post-hoc interval $[-.184873, .035713]$. Both intervals use 20,000 seed-stratified bootstrap draws; the diagnostic bootstrap seed is 20260906.

Every certificate at the 40-step probe equals one for all 130 development and 312 held-out candidates. Its standardized feature is consequently zero. The shadow-failure feature is also identically zero. A few earlier-checkpoint bounds are below one, so this observation is specific to the selector probe. The prospective cohort retains aggregate traces, while per-example probabilities needed for alternative margin thresholds are absent.

Development-only attribution checks. We evaluate six fixed feature variants on the ten original development blocks, using the same five ridge penalties, within-block standardization, and leave-one-block-out selection rule. These exploratory fits were chosen after the original decision and use development rows for both fitting and evaluation. A smaller CV regret is better. The certificate and failure columns are constant, so the additive-response interaction row reproduces the original full model's CV decisions.

Table 11 isolates the contribution of each added feature family on the development blocks.

Table 11: Exploratory feature ablations. "Response" in the additive rows is $\log(1 + \max_k D_{\mathrm{sym,k}}/\delta^2)$; fixed-relative KL uses $\log(1 + \max_k D_{\mathrm{sym,k}})$. All rows include rate and probe features.

Added features	Ridge penalty	CV mean regret
None	100	.044222
Margin	100	.044222
Additive response	.1	.054517
Fixed-relative KL	100	.044222
Margin, additive response, interaction	.1	.054517
Margin, fixed-relative KL, interaction	1	.054517

The rate-plus-probe baseline has the lowest CV regret on these ten development blocks; every added feature family ties or worsens it. The joint response, margin, and interaction variant also mixes their effects. This small development study offers no signal for advancing that feature set.

Compute and horizon. The 40- and 120-update horizons equal 1.25 and 3.75 passes over the training subset. A deployable 13-rate probe search with continuation costs at least $13 \cdot 40 + 80 = 600$ SGD updates. Stopping each shadow at the probe still gives $13 \cdot 80 + 80 = 1120$ updates for the response policy, before extra prediction evaluations. An 80-step probe-only comparator would cost $13 \cdot 80 + 40 = 1080$ updates. The archived producer executes 3120 updates per block because all bases and shadows reach 120; 1040 of those are shadow updates after the last response checkpoint. Wall-clock comparisons would additionally require timed implementations. The loss cap at 20 is inactive in every saved candidate.

G.4 Conditional block uncertainty for the breadth audit

Eligibility and scope. The breadth roster uses the archived functional-selector collection. Of 287 candidate rows, 276 survive the complete declared probe and have finite positive susceptibility. The 11 excluded rows fail at steps 5–38 in the 40-step families; positive partial-probe scores are ineligible. Every one of the 19 retained curves has at least one strict reversal and a changed maximizing rate. This collection remains separate from the 24-curve post-hoc prospective audit.

Each curve is a dataset/model/optimizer/seed combination in `functional_selector_runs.csv`. Ties are excluded from strict reversals. The point estimates are the mean of 19 curve reversal fractions (36.46977%) and the pooled numerator divided by pooled denominator (639/1929, 33.12597%). The resampling units are shared dataset/seed blocks, stratified as Fashion-MNIST seeds 0–4; MNIST seeds 0–2 with all three optimizers jointly; MNIST seeds 3–4 with SGD only; and SVHN seeds 0–2. Within each stratum, blocks are sampled with replacement while maintaining the original number of blocks. This preserves the fixed group roster and incomplete optimizer-availability pattern.

Table 12 gives the curve counts and strict reversal fractions for the declared breadth collection.

Exact enumeration of 37,800 weighted multinomial-count compositions represents 9,112,500 ordered bootstrap resamples. Weighted inverse-CDF 2.5% and 97.5% quantiles give $[32.71058\%, 40.40365\%]$ for the equal-curve mean and $[29.68345\%, 37.00418\%]$ for the pooled fraction. These intervals condition on the recorded dataset, optimizer, and grid roster. Each stratum has two to five seed blocks, which limits their resolution; variation across datasets or experiment designs lies outside this resampling calculation.

Table 12: Finite-shadow coordinate reversals in the declared functional-selector collection. Fractions are pooled within each row. All 19 curves have at least one reversal.

Dataset	Optimizer	Curves	Reversed pairs	Fraction
Fashion-MNIST	SGD	5	198/617	32.1%
MNIST	SGD	5	166/632	26.3%
MNIST	momentum	3	101/347	29.1%
MNIST	Adam	3	94/198	47.5%
SVHN	momentum	3	80/135	59.3%
All	Pooled	19	639/1929	33.1%

H FROZEN BRANCH-INTERVENTION DEVELOPMENT PILOT

Table 13 summarizes the frozen pilot's forecast, refinement, and chosen-anchor outcomes before the complete endpoint-level account.

Table 13: Frozen development pilot. Discrepancy is $d(a,b) = |a-b|/\max(a,b,10^{-10})$. Forecast medians use 24 correlated anchor/law rows per regime; refinement medians use two first-seed reference-anchor comparisons. Their thresholds are 50% and 25%. Response means average the chosen anchors across seeds, stages, and laws; all entries are descriptive.

Regime	Forecast discrepancy	Refinement change	Mean chosen-anchor response		
			Matched	Clean loss	Disagreement
MNIST / SGD	90.16%	83.26%	.000161	.000161	.000161
SVHN / momentum	93.19%	64.56%	.037412	.040530	.037339

Pilot design and data. The branch-pilot design was fixed before its smoke test and pilot outcomes, after the earlier prospective selector had failed its continuation rule. It draws from official training data. MNIST has 4,096 fitting examples and SVHN 8,192; each has disjoint 512-example clean-quality, 256-example calibration, and 512-example audit sets, plus a fixed 32-example activation diagnostic subset of calibration inputs. Fixed split seeds are 62001 and 62002. Pilot seeds 6201 and 6202 are separate from smoke seed 6200. The replication units are four regime/seed trajectories, each observed at two correlated stages under fixed datasets, splits, architectures, and recipes. Seed identifiers are reused across regimes; interpretation and continuation gates are regime-specific.

The MNIST CNN has convolutional widths 8 and 16, a 48-unit hidden layer, ReLU activations and max pooling. The SVHN residual CNN has widths 16/32/64, GroupNorm, and ReLU. There is no dropout, augmentation, or BatchNorm running-state estimation. Reference rates are .12 and .008; MNIST uses momentum zero and weight decay zero, while SVHN uses momentum .9 and weight decay .0005. Both use batch 128 and a declared 2,000-update constant-rate schedule. Only the required prefix through 1,000 updates is executed, retaining checkpoints at 200 and 1,000. Both pilot prefixes in each regime must meet the pre-target 1,000-update quality gate: accuracy at least .90 on MNIST and .60 on SVHN. All four pass before target branches are generated. Early SVHN branch endpoints have 19.5–43.9% quality accuracy, reflecting the different learning stages represented by the two origins.

Every stage branches from the same model and optimizer state at anchors $\eta \in \{.5, 1, 2\}\eta_{\mathrm{ref}}$ for 128 updates. The branch origin is fixed independently of the intervened branch rate; this differs from rerunning the whole prefix under that rate. Momentum buffers are deep-copied before restoration, and the branch learning rate is set after loading optimizer state. Complete minibatch permutations carry unused tails across epochs. The saved stream contains all post-checkpoint training randomness in these architectures; CPU and MPS RNG states are retained as well. Training permanently stops on any nonfinite loss, gradient, or parameter, without skipped updates or rescue. The real-data smoke passes in both regimes; equal-rate copies have exactly equal saved logits. All eight pilot stages complete.

Measurements and controls. Let D denote half of forward plus reverse predictive KL. Calibration at an anchor uses $h = .005\eta$ and the two-direction finite coefficient

$$\widehat{c}(\eta) = \frac{D(\eta, \eta + h) + D(\eta, \eta - h)}{2h^2}.$$

When the local expansion applies, this coefficient approaches the terminal KL coefficient $q_H^\eta = \frac{1}{2} s_H^\top F_Q s_H$, where $s_H = \partial_\eta \theta_H$ is the branch-end parameter tangent. The resulting forecast is $\widehat{c}(\eta)\mathbb{E}[\Delta\eta^2]$. For seed 6201 at both branch origins, the reference anchor also uses steps $.0025\eta$ and $.01\eta$. These compare finite coefficients on the same calibration inputs in the nonsmooth networks.

For each regime/seed/stage, four independently generated $U \sim \mathrm{Unif}[-.05, .05]$ draws are used under both laws and all anchors. The absolute law has $\Delta\eta = \eta_{\mathrm{ref}} U$; the relative law has $\Delta\eta = \eta U$. Thus the absolute law can be a 10% relative change at the smallest anchor. The laws coincide physically at the reference anchor and reuse those same branches. A branch is reused only when its source-checkpoint hash, rate, and stream hash match; reuse is explicit in the ledger. We report the analytic uniform second moment for anchor choice, and the realized four-draw moment for the numerical forecast diagnostic. The latter removes random second-moment variation, but retains finite-response approximation and independent calibration/audit sample differences. The primary target is finite symmetric KL on audit inputs; disagreement and signed error change are retained as secondary measurements.

Clean-quality loss limits eligible anchors to at most 1.05 times the best base-branch loss. Eligibility is determined without audit responses. There are three eligible-choice blocks out of four on MNIST and four out of four on SVHN, with both seeds represented; the single-eligible MNIST case remains in every result. The candidate rules use law-matched coefficient, unscaled additive coefficient, raw calibration KL, direct calibration disagreement, direct calibration TV, clean loss, probability margin, and recent base training loss. Exact ties choose the smaller rate. A raw-KL rule with the correct moment correction coincides algebraically with the law-matched rule. All candidate rules are specified directly from measured features.

Frozen continuation decision. All branches must complete and both regimes must pass the eligibility and instrument gates. The instrument requires median realized-moment forecast discrepancy at most .5 and median .0025-versus-.005 coefficient discrepancy at most .25, using $d(a, b) = |a - b| / \max(a, b, 10^{-10})$. Both values below 10^{-10} define an unresolved null case; none occur in the 48 endpoint forecasts. These empirical development screens govern continuation, which requires either a repeated supported law reversal in each regime or at least 5% lower mean response than both clean loss and direct disagreement in each regime, with a favorable contrast for each seed. Practical floors of 10^{-4} nats for the control response and 10^{-5} nats for the paired decrease prevent a percentage improvement on negligible responses from authorizing confirmation.

A supported reversal requires opposite anchor choices across laws, opposite observed response preferences with at least 75% paired-draw sign agreement and 10% relative separation, and a second seed reproducing the same pair's directions at the same checkpoint. The sole changed law-dependent choice, SVHN seed 6201 at step 200, fails the absolute-law direction. There are zero supported reversals. MNIST matches clean loss and direct disagreement exactly. SVHN's matched-minus-clean contrast is $-.009536$ for seed 6201 and $+.003300$ for seed 6202; its matched-minus-direct-disagreement contrast is $+.000073$ overall. The latter control is slightly better, and the clean-loss comparison fails the per-seed criterion. The frozen rule therefore stops the study at the pilot stage.

Figure 4 displays every observed-versus-predicted endpoint pair in the frozen branch pilot.

Execution and independent verification. The pilot schedules 280 branch requests and executes 248 unique physical branches after 32 recorded reuses. It performs 31,744 branch updates plus 4,000 prefix updates, 1,189,632 prediction examples, and 583.2 summed seconds measured inside prefix and branch routines on the local MPS device. These counts exclude setup, analysis, and the separate real-data smoke, which performs 1,424 updates. Prediction counts include repeated evaluations and activation diagnostics; they are not counts of unique examples. Every training curve starts at update zero, and saved predictions cover branch steps $0, 1, 2, 5, 10, 20, 40, 80, 128$.

The first saved-data analysis stops while serializing a NumPy Boolean to JSON, after generating per-example arrays. Its traceback and partial output are retained. A separately recorded wrapper

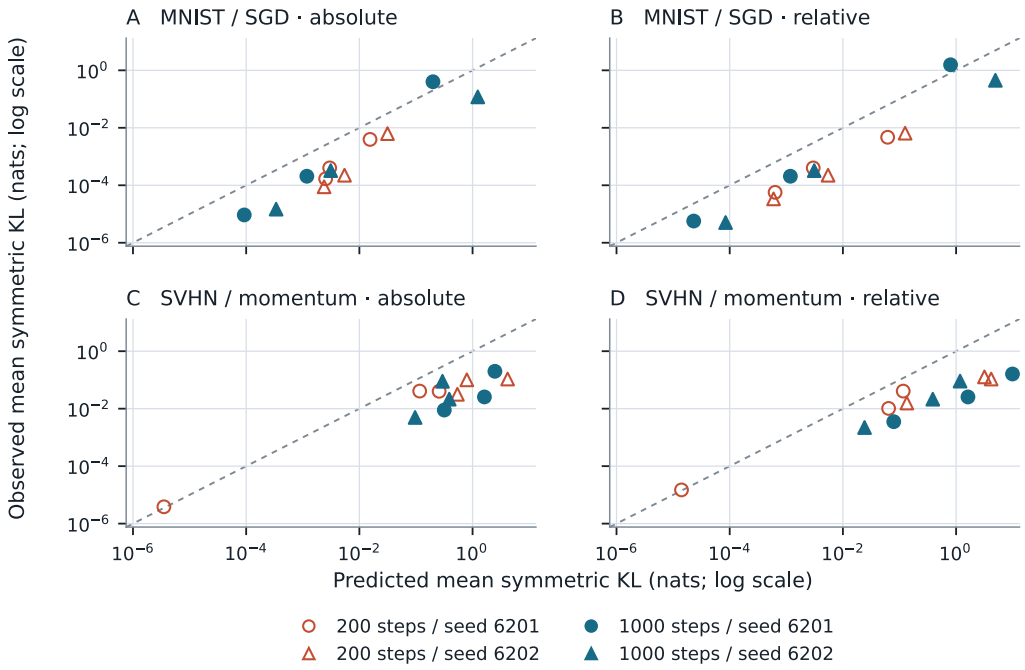


Figure 4: All 48 endpoint forecasts in the frozen branch pilot, twelve per panel. Each symbol is one seed/stage/anchor/law row; open orange and filled blue denote steps 200 and 1,000, and shape identifies seed. The dashed diagonal is equality and both axes are logarithmic. Predictions use the realized four-draw second moment and calibration inputs disjoint from the audit inputs. Discrepancy is $|a - b|/\max(a, b, 10^{-10})$. The 48 rows come from four correlated trajectories; the display is descriptive.

converts NumPy scalars to equal-valued Python scalars only at JSON output and replays the same analysis. The seven pre-outcome frozen files remain byte-identical, and no training is repeated. An independent reconstruction checks saved probabilities against raw logits, all 256 response pairs at 2,304 paired checkpoints (including recorded physical reuse), artifact hashes, state/stream identity, and the reported choices. A second review independently reconstructs endpoint coefficients and decision gates. These checks verify consistency of the saved experiment. The full companion package retains all branch artifacts; the manuscript package includes a portable, independently checkable endpoint probability snapshot.

I Post-hoc common-input endpoint diagnostics

This analysis was designed after the frozen branch pilot failed. It reuses the 512 quality examples per regime, which already determined clean-quality eligibility, and recomputes diagnostics from the saved branches. All eight stage blocks, all three anchors, both intervention laws, and all 248 physical branches remain in the source record. Forecast rows retain physical reuse at the reference anchor, giving 24 anchor/law rows per regime. For signed-loss counting, repeated physical base/target pairs shared by the two laws are removed, yielding 80 pairs per regime and 160 overall. Four regime/seed trajectories at two correlated stages supply these pairs; seed identifiers are reused across regimes.

Table 14 summarizes the exposed forecast and refinement diagnostics on the common quality inputs.

Let $h_c = .005\eta$ and let $\overline{D}_c$ be the mean symmetric KL from the two calibration branches at $\eta \pm h_c$, now evaluated on the common quality set. For the original four target increments $\Delta\eta_j$ in each row,

Table 14: Exposed development diagnostics on common quality inputs. Forecast medians use 24 correlated anchor/law rows per regime; refinement medians use the two first-seed reference-anchor origins. Discrepancy is $|a-b|/\max(a,b,10^{-10})$. Constant and linear controls reuse the same calibration branches with fixed exponents and no fitted parameters.

Diagnostic	MNIST / SGD	SVHN / momentum
Original / common-input quadratic forecast	90.16% / 90.80%	93.19% / 93.21%
Original / common-input refinement	83.26% / 81.62%	64.56% / 67.79%
Common-input constant forecast	70.05%	36.23%
Common-input linear forecast	44.45%	67.30%

the unfitted controls are

$$\widehat{R}_s = \overline{D}_c \frac{1}{4} \sum_{j=1}^{4} \left(\frac{|\Delta\eta_j|}{h_c} \right)^s, \qquad s \in \{0,1,2\}. \tag{25}$$

All realized target increments are nonzero. The quadratic rule reproduces the common-input second-moment forecast. The constant rule represents a finite-scale plateau. The three exponents were chosen after observing failure and applied without fitting target responses. Their regime-wise winners differ. Table 15 shows both seeds, including the marked SVHN seed difference.

Table 15: Common-input forecast median discrepancy by training seed (%). Each cell retains 12 correlated stage/anchor/law rows. The final column counts mean-loss decreases among 40 distinct target pairs per seed. Entries describe two seeds per regime.

Regime	Seed	Constant	Linear	Quadratic	Loss decreases
MNIST / SGD	6201	73.02	33.67	90.66	16/40
MNIST / SGD	6202	58.27	49.63	94.69	18/40
SVHN / momentum	6201	80.23	67.30	87.81	20/40
SVHN / momentum	6202	17.00	66.96	93.43	26/40

For each distinct base/target pair, we average signed alignment $(p_i - e_{y_i})^\top(\widetilde{f}_i - f_i)$, forward KL, and cross-entropy change over the same 512 labeled inputs. In the alignment-dominance counts, the absolute value is taken after averaging the signed term. All 80 pairs per regime have nonzero mean loss change; there are 46 increases on MNIST and 34 on SVHN. The MNIST mean-loss changes range from $-.015108$ to 2.129140 nats, and the SVHN range is $[-.166271, .104153]$. Every anchor is retained, including anchors that were ineligible for the frozen policy. The identity decomposes each observed loss change into its signed alignment and forward-KL components.

Horizon and input distinctions. The original calibration inputs have saved predictions at multiple branch horizons. Table 16 compares coefficients at relative increments $.0025$ and $.005$, at the reference anchor and the two origins of seed 6201. Its median is over two dependent stages per regime. The common quality inputs are saved only at the endpoint, so this horizon table is a separate calibration-input diagnostic. At one update the finite coefficients are close, whereas later discrepancies are substantial and nonmonotone. The table records horizon dependence; the route intervention below examines one possible source of short-horizon discrepancy.

Table 16: Original calibration-input refinement discrepancy by branch horizon (%). These post-hoc two-origin medians differ from the common-quality endpoint refinement in Table 14; all checkpoints use the original discrepancy definition.

Regime	1	2	5	10	20	40	80	128
MNIST / SGD	.0057	47.08	72.11	47.85	73.53	59.31	52.99	83.26
SVHN / momentum	.0100	11.97	7.06	19.59	33.86	57.03	79.19	64.56

Provenance and verification. Each consumed raw-logit file was checked against its saved branch ledger. Reconstruction of the original cross-input forecasts differs from the frozen analyzer by at most 1.09×10^{-15}; reconstructed quality losses differ from branch metadata by at most 4.45×10^{-16}. A separate implementation recovers the common-input forecast and refinement medians. The maximum per-example signed-loss identity residual is 3.60×10^{-14}. The bundled CSV files retain every forecast, refinement, signed-loss, and fixed-exponent row, with the original input manifest and independent check. The recorded pilot and selector decisions remain tied to their original protocols.

J FIVE-UPDATE TRAINING-ROUTE PROFILE

Question, origins, and intervention. The profile tests whether allowing the training route to change contributes to finite-response discrepancy over five updates from a fixed saved origin. Its existing checkpoints and quality inputs are exposed development evidence. The initial block is MNIST/SGD seed 6201 at prefix step 200. The frozen continuation roster comprises seeds 6201 and 6202 at origins 200 and 1,000 in each of the MNIST/SGD and SVHN/momentum regimes: four regime/seed trajectories with two correlated stages each. Seed identifiers recur across regimes. The study reuses the saved data split and training streams, with trajectories as the replication units.

Parameters, optimizer buffers, training minibatches, and the 512 quality-example IDs come from the archived branch-pilot origin. The intervened prefix is held fixed, so its rate derivative is zero. Rates and optimizer settings are loaded from the original configuration and checkpoint with equality checks. For each origin, the roster uses $\eta' = (1+\epsilon)\eta$ with $\epsilon \in \{0, \pm.0005, \pm.001, \pm.0025, \pm.005, \pm.01\}$, two training conditions, CPU float32 and float64, and five updates. State is retained after every update; predictions and tangents are reported at $H = 0, 1, 2, 5$.

Ordinary paths use natural training routes. The implemented controlled condition applies the base-rate trajectory's recorded ReLU gates and max-pool indices at the corresponding training update. It also records which natural routes the controlled state would have used. In V2, these base routes are captured in a separate no-grad forward on detached parameters; the reference value gradient uses fresh gradient-enabled leaves. The postmortem in Appendix L shows that these two capture contexts can record different routes even when value paths satisfy preflight tolerance. All endpoints use the original inference function with natural routes. This development intervention motivated the later recorder that captures routes inside the actual gradient forward.

Optimizer-state AD reference. Write g_k for the minibatch loss gradient, μ for momentum, and λ for weight decay. For the saved non-Nesterov, zero-dampening momentum state, the recurrence is

$$b_{k+1} = \mu b_k + g_k + \lambda\theta_k, \qquad \theta_{k+1} = \theta_k - \eta b_{k+1}.$$

With $s_0 = \dot{b}_0 = 0$, its declared smooth AD continuation propagates

$$\dot{b}_{k+1} = \mu\dot{b}_k + (H_k + \lambda I)s_k, \qquad s_{k+1} = s_k - b_{k+1} - \eta\dot{b}_{k+1}.$$

For plain SGD, use $\mu = 0$ and the corresponding gradient direction without persistent buffer state. Hessian-vector products propagate this reference in each training condition; the output JVP uses the unchanged inference model. For logits f_i and output tangent $\dot{f}_i$ on $n = 512$ quality examples, the reported coefficient is

$$q_H^{\mathrm{AD}} = \frac{1}{2n}\sum_{i,c} p_{ic}\left(\dot{f}_{ic} - \sum_j p_{ij}\dot{f}_{ij}\right)^2, \qquad p_i = \mathrm{softmax}(f_i). \tag{26}$$

This is the AD reference under the recorded conventions. Exact zero preactivations and positive max-pool ties are retained on training and quality inputs, where the implementation selects its declared branch convention.

Preflight and numerical resolution. Before each perturbation roster, five base-rate updates compare native `torch.optim.SGD`, the functional ordinary path, the path with fixed training routes, and an equal-rate duplicate in each precision. The maximum parameter and buffer differences at every step must satisfy the frozen precision-scaled tolerance. Tangent primal values must also agree

with the corresponding ordinary or controlled base. Synthetic checks compare AD with a reverse unroll and a smooth finite difference, including momentum and weight decay. These checks validate the implemented smooth-reference calculation against the specified base paths.

At each recorded base checkpoint, let e_f be the larger of the equal-copy maximum logit difference and $64\varepsilon_{\mathrm{mach}}(1+\|f\|_\infty)$. With $C = 10$ classes and null-copy KL R_0, the declared screening floor is

$$\tau_H = \max\{10^{-14}, 100R_0, (C/4)e_f^2\}.$$

This numerical screening floor determines resolution: a case is resolved when $\max\{R_H, q_H^{\mathrm{AD}}\delta^2\} > \tau_H$, and its discrepancy is

$$d_H = \frac{|R_H - q_H^{\mathrm{AD}}\delta^2|}{\max\{R_H, q_H^{\mathrm{AD}}\delta^2, \tau_H\}}. \tag{27}$$

Both precisions and all unresolved cases remain reported. The primary screen uses the 20 float64 sign/size cases at $H = 2, 5$, requiring at least 16 to be resolved in both conditions. Both medians use the identical jointly resolved subset. Continuation requires an ordinary median at least .1 and a reduction of at least 50%, together with base and derivative correctness. These development screens were frozen before the initial perturbation outcomes. The initial pass authorized the unchanged eight-block roster.

Complete paired results. All eight declared blocks complete with 20/20 jointly resolved primary cases (Table 17). The reported ordinary and fixed-route AD coefficients are identical at $H = 2, 5$ within every block; the response changes are therefore compared with the same quadratic reference. The initial block passes the frozen screen with 98.03% relative reduction of the median; 16/20 individual paired discrepancies improve. All four SVHN origins also exhibit material ordinary discrepancy and 97.79–99.60% median reduction, with all 20 individual cases improving in each block. The other three MNIST ordinary medians are below .01. Every scheduled origin is retained, including these small-error cases. The comparison describes four training trajectories at two dependent stages each; the later prospective study supplies the independent trajectory-level test.

Table 17: Complete paired training-route profile. Median discrepancy follows Eq. (27) on 20 paired float64 cases per origin (two signs, five sizes, $H = 2, 5$). Joint 64/32 gives resolved counts by precision; displayed medians use float64. The first row is the predeclared screen and the remaining rows are descriptive.

Regime	Seed/origin	Ordinary	Fixed routes	Reduction	Joint 64	Joint 32
MNIST	6201/200	.271360	.005340	98.03%	20/20	20/20
MNIST	6201/1000	.002988	.001940	35.07%	20/20	8/20
MNIST	6202/200	.009939	.004421	55.52%	20/20	18/20
MNIST	6202/1000	.009359	.007502	19.85%	20/20	20/20
SVHN	6201/200	.226854	.000917	99.60%	20/20	20/20
SVHN	6201/1000	.179651	.000834	99.54%	20/20	20/20
SVHN	6202/200	.369455	.008161	97.79%	20/20	20/20
SVHN	6202/1000	.514728	.002246	99.56%	20/20	18/20

Float32 jointly resolves 8–20 primary cases across origins, with 8/20 at MNIST 6201/1000. Precision also changes some resolved medians: SVHN 6201/200, for example, has ordinary median .302169 in float32 versus .226854 in float64. Full precision-specific rows retain the resolved and unresolved cases.

Preserved V1 failure and V2 repair. The first SVHN V1 correctness preflight stops before any SVHN rate-perturbation endpoint. Its ordinary update used `torch.func.grad` values. A float32 discrepancy from native SGD grows to a 9.6241×10^{-5} momentum-buffer difference by update two, exceeding the frozen tolerance. An independent base-only diagnosis finds an initial gradient discrepancy of 8.1956×10^{-8} and three changed ReLU decisions on the next minibatch. The failed record, source, and tolerances are retained unchanged.

The frozen V2 successor uses ordinary `torch.autograd.grad` value gradients with `create_graph=False`. It discards the primal gradient returned by the `torch.func` HVP calculation and uses that calculation only for the derivative reference, adding one charged ordinary

gradient evaluation per tangent update. This defines a hybrid smooth AD reference with native value gradients and transform-based HVPs under the recorded tie convention. Base-only source-bound repair checks recover exact parameter and buffer equality with native SGD at the previously failing origin in both precisions.

The rate grid, horizons, tolerance, resolution rule, initial screen, and roster are unchanged. V2 admits only the four previously declared SVHN blocks and verifies the sealed V1 initial pass in its original runtime module namespace before each phase. The completed MNIST perturbation roster is not rerun. This is an implementation repair before the affected scientific outcomes, with the V1 failure retained as evidence.

Simple controls. For the larger target sizes $.0025, .005, .01$, the same pair of calibration branches at relative size $.001$ supports three unfitted powers, $s = 0, 1, 2$, and a sign-matched finite-logit secant. For the power controls, multiply the mean calibration KL by $(|\delta|/\delta_{\mathrm{cal}})^s$. The secant extrapolates logits from the calibration endpoint of the target sign. Conditional quadratic and linearized-logit AD forecasts use $q_H^{\mathrm{AD}}\delta^2$ and the KL to $f_H + \delta\dot{f}_H$, respectively. The final control uses a quadratic coefficient from the finer calibration at $.0005$; the dyadic check charges four calibration branches when both levels are used.

Table 18 uses every method on the same 12 float64 ordinary-training target cases per origin: three larger sizes, both signs, and $H = 2, 5$. Its bounded discrepancy uses the observed KL, the method's forecast, and the same case-specific numerical floor in the denominator. Coefficients come from calibration branches, while all controls share the same target inputs, physical rate, and horizon. At SVHN 6202/1000, for example, the coarse quadratic median is $.348$, versus $.382$ for conditional AD quadratic. These target-case medians use a different subset from the 20-case primary route comparison.

Table 18: Median bounded discrepancy for every simple control, using the same 12 ordinary-training float64 cases per origin: two signs, three relative sizes, and $H = 2, 5$. Fixed-power, matching-sign secant, conditional-AD, and finer-calibration forecasts use the definitions above. These are exposed development comparisons.

Regime	Seed/origin	S_0	S_1	S_2	Secant	AD quad.	AD logits	Dyadic
MNIST	6201/200	.783	.746	.64	.58	.335	.334	.877
MNIST	6201/1000	.974	.85	.00847	.00508	.00783	.00703	.00783
MNIST	6202/200	.963	.83	.0547	.0269	.0186	.0184	.0188
MNIST	6202/1000	.96	.8	.0295	.0221	.0111	.0181	.0158
SVHN	6201/200	.979	.892	.433	.603	.192	.19	.496
SVHN	6201/1000	.936	.653	.36	.357	.18	.18	.465
SVHN	6202/200	.925	.623	.488	.406	.246	.246	.627
SVHN	6202/1000	.914	.528	.348	.472	.382	.382	.573

Cost and provenance. Each complete block has 44 five-update paths (220 optimizer updates), 40 preflight base updates, and 20 tangent-primal updates/HVPs. Across eight blocks, the ledger records 1,760 profile optimizer updates, 320 preflight updates, 160 tangent-primal updates, and 160 HVPs. V2 adds 80 ordinary value-gradient evaluations for its tangent paths. Three recorded root base-only repair checks add 60 updates, and independent base checks add 40 updates and 10 HVPs. Synthetic tests, the stopped V1 preflight, output derivatives, inference, and other independent diagnostics add further costs. This ledger accounts for the development comparison's measured work.

The final summaries derive from 1,280 response rows and 4,032 control rows, retaining both precisions and every case. Each source and row file is checked against its sealed completion record. The portable snapshot identifies the V1 and V2 blocks and keeps the failed V1 preflight separate from completed V2 outcomes. Available CPU/MPS comparisons at matching saved checkpoints are recorded separately from the common-backend precision intervention. The next section gives the independent successor study.

K PROSPECTIVE ROUTE CONFIRMATION

Restart disclosure and frozen design. The interrupted predecessor's 13/16 descriptive SVHN successes motivated a new, independently frozen trajectory roster. The event retains its .1 and 50% thresholds. Before any successor prefix or outcome was generated, the design fixed $n = 20$, a one-sided exact binomial test of $H_0 : p \leq .5$, and confirmation at 15 or more successes. The nominal level is .025 and the exact null size is .0206947327; planned power is .804208 at $p = .8$. The study tests the pattern seen in development and the interrupted campaign on new local trajectories.

The target distribution is conditional on the exposed SVHN data pool, TinyResNet architecture, CPU float32 prefix protocol, update-200 origin, and hash-based seed generator. Twenty distinct initialization seeds and 20 distinct minibatch-stream seeds were generated from a new namespace. The local audit finds no collision with available development or interrupted-campaign seeds. Each prefix logs checkpoints from step zero and the full 200-update loss curve; trajectory selection is fixed independently of accuracy.

Successor intervention and complete event. The five-update profile uses the same five minibatches, natural inference, both CPU precisions, and the reference plus both signs of relative increments $.0005, .001, .0025, .005, .01$. Each trajectory therefore has 44 physical paths. Unlike the interrupted instrument, the successor condition records ReLU masks and max-pool indices inside the actual gradient-enabled reference forward, then replays those base-path routes only during training. The ordinary, native-SGD, and successor fixed-route base paths must be bitwise equal through all five updates in both precisions before the profile is eligible. Synthetic identities, source/input receipts, tangent-primal checks, numerical floors, and route-tie records are also retained.

Each trajectory contributes one binary observation. It succeeds only when all checks pass and, on the same float64 cases at $H = 2, 5$, at least 16 of 20 cases are resolved in both conditions, the ordinary median bounded discrepancy from the conditional AD quadratic reference is at least .1, and route replay reduces that paired median by at least 50%. The fixed denominator contains all 20 scheduled trajectories; cases failing any criterion count zero. Float32 repeats are descriptive.

Primary result. All 20 scheduled prefixes, dual-precision preflights, and profiles complete. Every preflight has exact base-path equality, and every trajectory has 20/20 jointly resolved primary cases. Fifteen satisfy the complete event, exactly reaching the predeclared boundary:

$$\begin{aligned} \widehat{p} &= 15/20 = .75, \\ p_{\text{exact}} &= \Pr\{\text{Bin}(20, .5) \geq 15\} = .0206947327, \\ \text{CI}_{.95} &= [.508954, .913429]. \end{aligned}$$

The interval is equal-tailed Clopper–Pearson. The same 15 rows meet the descriptive float32 event. Across trajectories, the ordinary median E is .199305, the fixed-route median is .002917, the median within-trajectory relative reduction is 97.99%, and the median fold ratio is $50.17\times$. Individual reductions range from 93.62% to 99.67%. Because correctness, resolution, and reduction pass for all 20, the five non-events are exactly the trajectories with ordinary median $E < .1$.

Table 19 reports all 20 frozen successor trajectories and their event status.

Absolute scale and matched controls. The event measures relative disagreement with a local quadratic forecast. For each trajectory, taking the median symmetric KL over the same 20 ordinary float64 cases gives a median of trajectory medians 2.4814×10^{-7} and range $[1.5467 \times 10^{-8}, 4.5795 \times 10^{-6}]$. Under route replay the corresponding median is 2.2768×10^{-7} and range $[1.5649 \times 10^{-8}, 4.5693 \times 10^{-6}]$. These values show the small absolute scale on which the relative forecast improvement occurs.

Table 1 uses all seven inherited forecasts on the same 12 coarser float64 cases per condition and trajectory. Conditional AD is competitive in ordinary training, but under route replay the matching-sign logit secant has the smallest median across trajectories. The fixed quadratic, fine-grid quadratic, and conditional AD values are nearly identical under route replay. These controls show a route-sensitive pattern across the quadratic-family and secant forecasts.

Independent verification and retained records. After the one authorized formal run, a raw replay recomputes the saved outcomes. A separate zero-update verifier regenerates all 20 trajectory events from CSV, the seed lineage, exact p-value and interval, and all source/input/receipt/raw-artifact hashes. It reports 20 unique initial states, 20 unique streams, 15 successes, and `PASS`. The exact SHA-256 values are:

Frozen design	`7bc42388078a9045211f890c95b53bc5`
	`0bbcaf49414c159926b5f0e0a34fa1ca`
Summary	`89813e13f853c76d427b427e11473c72`
	`87092893c1d330ddbdad36299da459ca`
Verification	`7eda2337f517db15b8d4e19eee39895f`
	`f3cf7e8af31ec7338dc91e9d82f34b0e`

The manuscript package retains the protocol, frozen configuration and receipt, complete trajectory CSV, final summary and verifier record, analysis source and tests, and the derived figure/control tables. The full tensor artifacts remain in the separate immutable experiment package. The inference uses the 20 successor trajectories.

Table 19: All 20 trajectories in the prospective route confirmation. Medians are over the 20 paired, resolved float64 cases at $H = 2, 5$. Reduction is $100(1 - E_{\text{fixed}}/E_{\text{ordinary}})$. The event also requires all correctness checks and at least 16 resolved cases; both conditions pass for every row.

Rep.	Init. seed	Stream seed	Ordinary E	Fixed E	Reduction	Event
1	325599416	160981301	0.115647	0.004044	96.50%	met
2	138230883	152149406	0.204407	0.004435	97.83%	met
3	984558111	1052190105	0.177200	0.001791	98.99%	met
4	183631589	1837938200	0.207854	0.000679	99.67%	met
5	1616635711	1619370413	0.219951	0.010451	95.25%	met
6	587694114	1040390443	0.242644	0.005564	97.71%	met
7	1514277370	1344188885	0.012159	0.000428	96.48%	not met
8	1172880120	1773818330	0.432454	0.013569	96.86%	met
9	426519036	706584922	0.213459	0.010767	94.96%	met
10	1946856480	681537716	0.042638	0.000217	99.49%	not met
11	2092194130	1668187443	0.029027	0.000275	99.05%	not met
12	303639858	459755535	0.200697	0.004879	97.57%	met
13	1665738916	754399159	0.201178	0.000724	99.64%	met
14	851238784	1621734040	0.275300	0.007815	97.16%	met
15	2092049494	518463450	0.093394	0.001722	98.16%	not met
16	431176723	1761105383	0.284804	0.004064	98.57%	met
17	968334476	1036810078	0.148707	0.001734	98.83%	met
18	1737446124	1828940155	0.071353	0.000978	98.63%	not met
19	316215514	1181150162	0.197913	0.012626	93.62%	met
20	574286734	928456801	0.110325	0.001501	98.64%	met

L INTERRUPTED PREDECESSOR ACROSS NEW TRAINING TRAJECTORIES

Declared population and roster. The predecessor protocol targets the complete training-route event under a fixed CPU-prefix distribution, conditional on the previously exposed data pools, architectures, and settings. Its roster specifies 16 new initialization/minibatch-stream pairs for SVHN/momentum, the sole primary regime, and 16 for descriptive MNIST/SGD analysis. Each trajectory contributes one origin at prefix update 200. Separate hash-derived initialization and stream seeds are treated as independent draws under the conventional pseudorandom-seed sampling model. The seed-exposure audit checks collisions within the available records. Prefixes use native SGD on CPU float32 with four threads and deterministic algorithms; earlier development prefixes used MPS, so this campaign also changes backend.

The original pilot configuration and tensor/index manifest specify normalization, learning rate, optimizer settings, batch size, and disjoint input partitions. The protocol retains the initial state and RNG, all 205 minibatches, training losses from step zero, and prefix checkpoints $0, 1, 2, 5, 10, 20, 40, 80, 128, 200$ with quality logits. Accuracy is recorded descriptively; the roster and design were fixed before trajectory outcomes.

Unchanged profile and complete event. From each saved origin, the five-update V2 profile in Appendix J uses the same five minibatches, both CPU precisions, two training conditions, and reference plus both signs of relative increments $.0005, .001, .0025, .005, .01$ (44 physical paths). Inference remains natural. The inherited base/optimizer/stream and tangent-primal checks, numerical floors, tie qualifications, and hybrid smooth AD reference remain unchanged.

A trajectory succeeds only if all correctness checks pass and its float64 $H = 2, 5$ profile meets all three criteria: at least 16 of the 20 sign/size/horizon cases are jointly resolved; the ordinary median of Eq. (27) is at least .1 on exactly those paired cases; and fixed training routes reduce that paired median by at least 50%. Too few resolved cases, small ordinary discrepancy, or insufficient reduction counts zero. The denominator is the full declared roster. All seven forecast controls and both precisions are descriptive.

Predeclared inference and failure rule. For SVHN, let p be the probability of this complete event under the declared CPU-prefix protocol. The predeclared test was one-sided exact binomial for $H_0 : p \leq .5$, with $n = 16$, $\alpha = .05$, and a critical count of 12 (null size .0384063721). The plan also prescribed an equal-tailed exact 95% Clopper–Pearson interval. The later instrument stop ends the campaign before this inference. MNIST remains descriptive.

The protocol orders all SVHN trajectories before all MNIST trajectories, without outcome-dependent stopping, expansion, replacement seeds, or changed thresholds. Genuine nonfinite prefix training, explicit nonfinite primal profile updates, or nonfinite ordinary primal predictions count zero. Tangent-only nonfinites and instrument, provenance, equality, unexpected-exception, or interruption failures halt the campaign and mark it inconclusive. The later successor was designed after this stop and retains the predecessor as a separate record.

Final interrupted-campaign status. Table 20 retains every scheduled trajectory. Sixteen SVHN and eight MNIST profiles complete; all 24 have 20/20 jointly resolved primary cases. The complete material event is observed in 13 SVHN profiles and two completed MNIST profiles. The other completed profiles remain in the descriptive record. The failed MNIST instrument check and seven unrun trajectories are unassessed for the material event.

Table 20: Final planned-roster accounting. Material events are descriptive counts among completed profiles. The campaign stopped at its instrument check, so planned primary inference is unavailable.

Regime	Planned	Completed	Events observed	Instrument failure	Unrun
SVHN / momentum	16	16	13	0	0
MNIST / SGD	16	8	2	1	7

The stop occurs at MNIST replicate nine, initialization seed 288406286, before that trajectory's perturbed profile. At the first float32 base update, the fixed-training-route state differs from native SGD

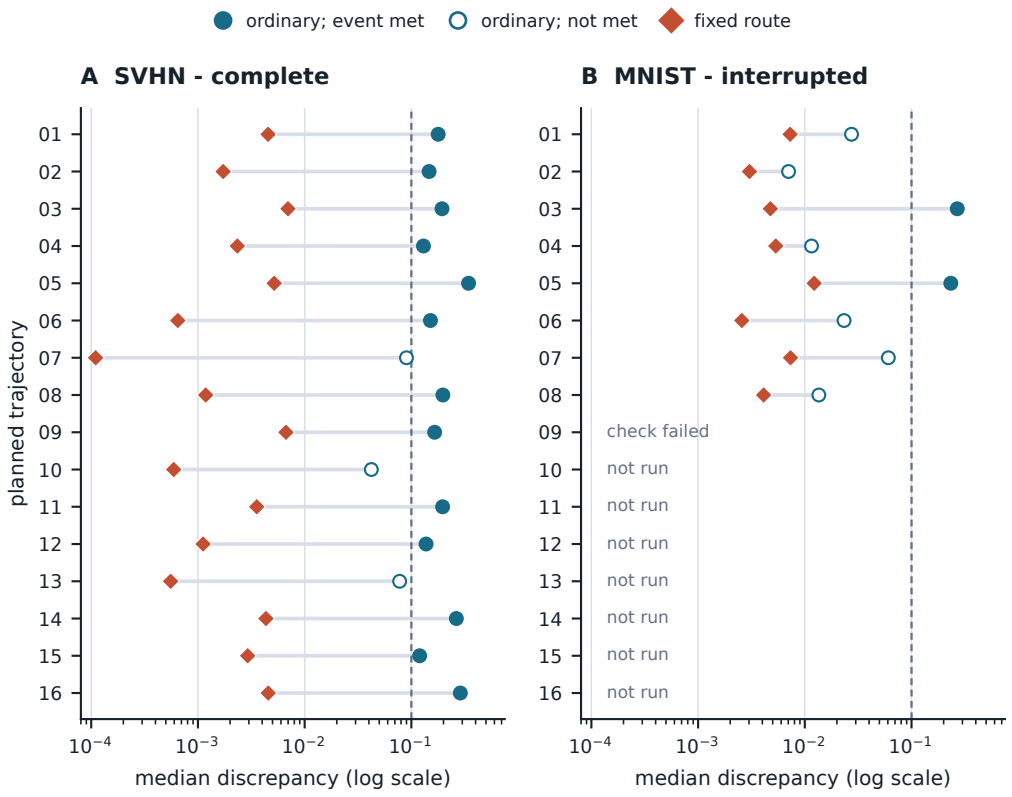


Figure 5: Descriptive profiles for all 32 planned trajectories, redrawn from the displayed six-decimal roster medians. Completed pairs use 20 float64 cases; axes are logarithmic and fixed route denotes the predecessor instrument. The .1 line marks the ordinary-discrepancy condition; filled or open ordinary circles show whether the full event was met. Unassessed rows remain visible. *SVHN: complete* labels the 16 completed SVHN entries; the campaign stopped at the MNIST instrument check.

by a maximum absolute parameter difference $2.078711986541748 \times 10^{-5}$, while ordinary/native difference is zero. This exceeds the frozen base-equality tolerance and triggers the instrument stop. The full campaign is marked `INCONCLUSIVE`, with a null primary p-value.

Descriptive profiles and controls. Table 21 and Figure 5 retain every planned row. Within all 24 completed profiles, ordinary and fixed-route reported float64 AD coefficients coincide, and all 20 primary cases are resolved in both conditions. Float32 resolves 16–20 joint cases depending on the trajectory; these counts and the unassessed rows remain in the roster. The descriptive medians coexist with individual case variation and the predeclared .1 ordinary-discrepancy threshold.

Absolute response scale. The event's .1 threshold concerns bounded relative discrepancy. Among the 16 completed SVHN trajectories, the ordinary within-trajectory median symmetric KL ranges from 2.23×10^{-8} to 3.66×10^{-6}; the median of these 16 medians is 2.85×10^{-7}, versus 2.63×10^{-7} for fixed training routes. Across the eight completed MNIST profiles, ordinary within-trajectory median KL ranges from 4.48×10^{-7} to 2.34×10^{-5}. The absolute KL scale is small across these descriptive trajectories.

Figure 6 compares all seven forecasts on the same 12 ordinary-training float64 target cases per completed trajectory: both signs of relative increments .0025, .005, .01 at $H = 2, 5$. Methods use the unchanged definitions in Appendix J. Displayed values are bounded discrepancy multiplied by 100, not absolute KL. The original study retained 672 summary cells across precisions and conditions;

Table 21: Every planned new trajectory at prefix 200. Medians use the 20 jointly resolved float64 primary cases and Eq. (27); J32 is the corresponding float32 count. Failed and unrun entries are unassessed; completed rows describe the stopped campaign.

Regime	Rep.	Init. seed	Ordinary	Fixed routes	J32	Event/status
SVHN	1	1191078306	.178347	.004545	20	met
SVHN	2	1569823481	.146626	.001723	20	met
SVHN	3	1666986268	.193655	.006981	18	met
SVHN	4	2101689349	.129860	.002339	20	met
SVHN	5	1371959192	.343651	.005182	20	met
SVHN	6	1897428980	.150956	.000648	20	met
SVHN	7	857041336	.089994	.000110	18	not met
SVHN	8	1926244198	.197342	.001183	20	met
SVHN	9	806720174	.165210	.006680	20	met
SVHN	10	1621971472	.042179	.000593	20	not met
SVHN	11	1730453352	.196199	.003560	20	met
SVHN	12	952991096	.137316	.001118	20	met
SVHN	13	1639616172	.077655	.000555	20	not met
SVHN	14	2147102331	.264012	.004337	20	met
SVHN	15	443527218	.119576	.002926	20	met
SVHN	16	806712418	.287727	.004560	20	met
MNIST	1	378787551	.027401	.007285	18	not met
MNIST	2	1059717778	.007042	.003030	16	not met
MNIST	3	789961011	.268336	.004757	20	met
MNIST	4	1336260133	.011564	.005346	20	not met
MNIST	5	154100488	.233436	.012235	16	met
MNIST	6	1747206946	.023321	.002568	18	not met
MNIST	7	1353866767	.060639	.007349	20	not met
MNIST	8	1220996994	.013536	.004114	16	not met
MNIST	9	288406286	—	—	—	check failed
MNIST	10	1532200734	—	—	—	not run
MNIST	11	182487255	—	—	—	not run
MNIST	12	485575830	—	—	—	not run
MNIST	13	156576466	—	—	—	not run
MNIST	14	1235647975	—	—	—	not run
MNIST	15	20622697	—	—	—	not run
MNIST	16	169182708	—	—	—	not run

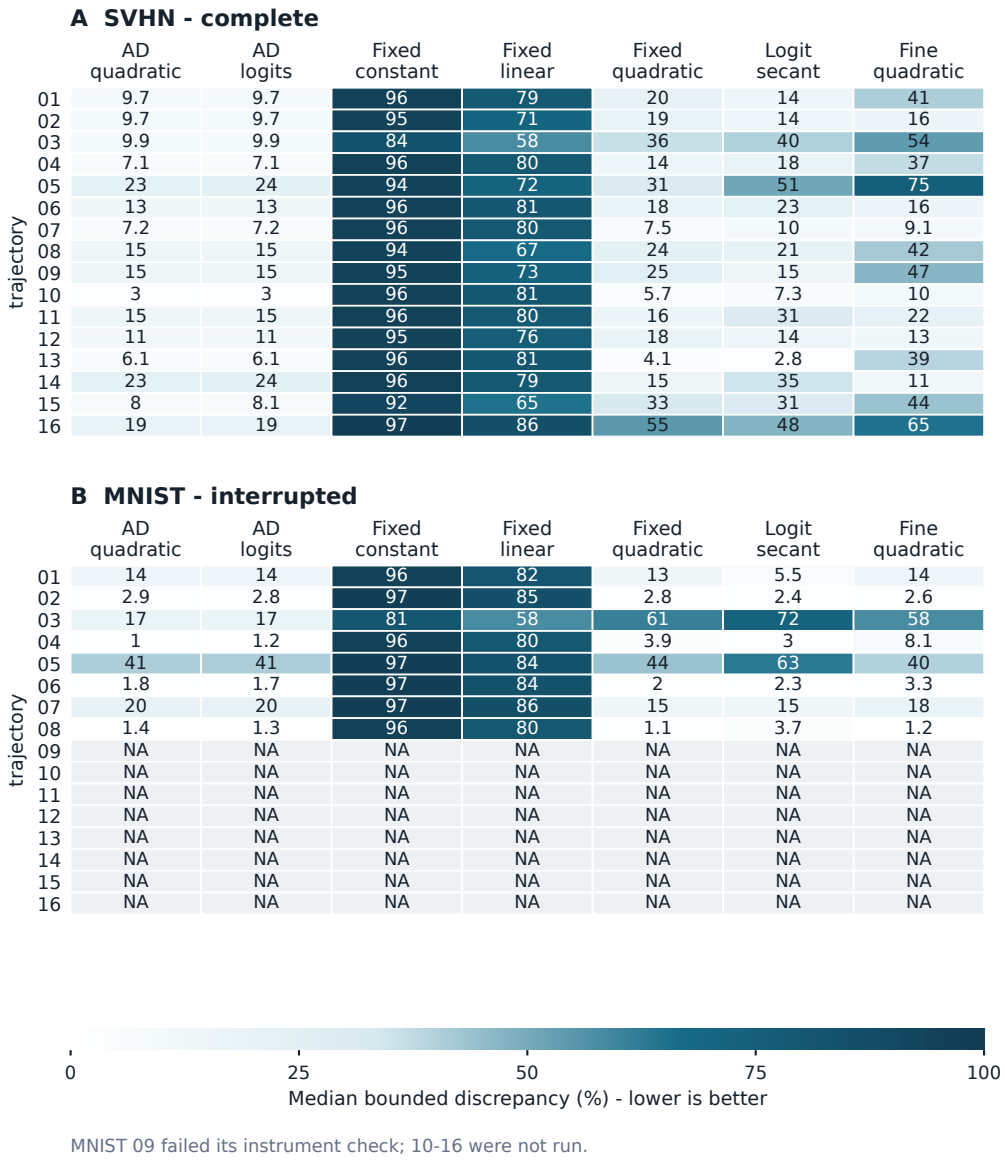


Figure 6: Seven matched controls on the planned roster, using the same 12 ordinary-float64 cases per observed trajectory. The heatmap reproduces the archived figure's displayed rounded values; color and labels show median bounded discrepancy $\times 100$, with lower values better. MNIST row 09 failed its instrument check and rows 10–16 were not run, so NA means unassessed. The comparisons summarize the completed profiles in the stopped campaign.

the heatmap shows the 168 available ordinary-float64 displayed values, plus explicit missing slots. This package retains those control values at the archived figure's rounding.

Recorded work. Completed and attempted phases record 5,000 prefix updates, 5,280 profile optimizer updates, 960 completed-preflight updates, 480 tangent-primal updates, 480 HVPs, and 480 additional reference-gradient evaluations. The failed preflight adds 20 float32 base updates: its four five-step paths were constructed before the equality loop stopped at update one. It executes no float64 preflight updates. Synthetic tests, inference, and the separate diagnosis add further work.

Bounded route-capture diagnosis. A separate postmortem isolates a mismatch between the route-capture and value-gradient execution contexts at the failed origin. The archived recorder uses detached parameters without gradient tracking in a separate no-grad forward; the actual reference gradient uses fresh gradient-enabled leaves. The two forwards differ by a maximum 1.9073486×10^{-6} at the first convolution, with 45,082 first-pool index differences, 3,075 later-pool index differences, and one ReLU-mask difference. These are counts of route record differences at the same origin.

Replaying the archived routes reproduces the $2.0787119865 \times 10^{-5}$ first-update mismatch with both the legacy gather implementation and a native-kernel replay. Using routes from the actual gradient forward removes that first-update discrepancy. At this origin, capture context explains the failed base-equality check. The earlier perturbed profiles remain development observations under their recorded instrument.

A capture-only successor records routes inside the actual reference-gradient forward, retaining the original ReLU multiplication and gather-based max-pool replay. Its initial check at the failed MNIST/SGD origin gives bitwise-equal native, ordinary, and successor fixed-route base paths at updates 0–5 in both precisions.

A separately frozen regression then uses every accessible CPU step-200 prefix: nine MNIST/SGD origins, including the failure, and all sixteen SVHN/momentum origins. A fixed two-origin smoke precedes one formal all-25 run. Each origin uses both precisions, the same five saved minibatches, and three base paths. All 900 parameter/buffer/logit pairing rows at states 0–5 are bitwise equal, with all three maximum differences exactly zero (Table 22). Thus the repair covers nonempty momentum buffers and the second architecture within this saved population.

Table 22: Frozen saved-origin base-path regression. "Exact" counts the three path pairings at six states per origin; the maximum triple is parameter/buffer/quality-logit absolute difference. Detached changes compare archived route capture with capture inside the gradient forward. Denominators are 68,014,080 entries/225 tensors per MNIST row and 1,006,632,960 entries/880 tensors per SVHN row; entries are correlated diagnostics.

Setting	Origins	Exact	Max. P/B/L	Detached changes
MNIST / SGD, float32	9	162/162	0/0/0	2,075,990 / 94
MNIST / SGD, float64	9	162/162	0/0/0	0 / 0
SVHN / momentum, float32	16	288/288	0/0/0	0 / 0
SVHN / momentum, float64	16	288/288	0/0/0	0 / 0

The archived capture differs at every MNIST/float32 origin, primarily in max-pool indices; saved MNIST/float64 and TinyResNet origins agree. The 60-update smoke and 750-update formal execution test base paths at saved origins, repeating the smoke origins within the full roster. A zero-update verifier independently recomputes both decisions. This base-path result supported the separately frozen successor design in Appendix K.

Independent artifact checks. A no-training audit checks all 24 completed analyses and their 1,056 saved physical paths, regenerates initialization and stream lineage for the 25 retained prefixes, and passes the frozen-source and raw-artifact integrity checks. The original metric verifier also recomputes the available raw outputs without errors or warnings. Its audit status is `INCOMPLETE`, reflecting seven missing scheduled receipts; the scientific campaign status is `INCONCLUSIVE`. The successor analyzes its own fresh trajectories. Exact source snapshots and check records accompany the revision.

M FROZEN SMOOTH-NETWORK CALIBRATION

This successor experiment is separate from archival training. It imports the unchanged `FlatMLP` and initializer: 12 tanh hidden units, 62 parameters, two-class softmax. Two-moons has 360 examples with noise .22; each of seeds 7, 17, 27 gives a stratified 241/119 train/validation split. The scaler is fitted on training examples. The older generator fitted on all 360 examples before splitting, so the two sweep designs are analyzed separately. The frozen design uses rates $.1, .5, 1.0$, relative steps $.005, .01, .025, .05$, 40 full-batch SGD updates, CPU float64, and one thread.

The protocol and runner are hashed before execution. A smoke test checks analytic scalar and softmax identities, tangent versus unrolled autodiff, logit Jacobian versus JVP, zero step, KL order, and the certificate (44 checks). One subsequent formal execution completes all nine bases and 36 shadows. There are no large-finite flags or failed runs. Every step 0–40 retains parameters, tangents, logits, probabilities, margins, per-example KL, disagreement, and status. The mean/maximum coefficient errors aggregate steps 1–40. Exact reference means range from .0075545 to 4.70587 and exact maxima from .124354 to 7.65595.

Table 23 gives the full seed/rate error ranges behind the main-text medians.

Table 23: Full calibration error ranges (%). Each cell is median [minimum, maximum] across the nine fixed seed/rate pairs.

ϵ	Forward mean	Symmetric mean	Forward maximum	Symmetric maximum
.005	.549 [.295, .643]	.594 [.394, .693]	.341 [.083, .448]	.440 [.120, .489]
.010	1.093 [.587, 1.280]	1.183 [.783, 1.378]	.681 [.167, .892]	.877 [.241, .975]
.025	2.698 [1.438, 3.147]	2.918 [1.921, 3.386]	1.690 [.419, 2.209]	2.174 [.606, 2.412]
.050	5.283 [2.785, 6.128]	5.706 [3.727, 6.586]	3.342 [.843, 4.348]	4.287 [1.223, 4.744]

The fixed margin grid is $\{.05, .1, .2, .3\}$. There are $36 \times 41 \times 4 = 5904$ margin rows, each evaluated under both KL conventions; at step zero the predictors agree. The 1440 positive-step comparisons include 57 with any disagreement, maximum .03361345 and mean .00053105. Minimum-grid bounds are nonvacuous in all 1440: the forward mean/maximum are .03784918/.72790904 and symmetric mean/maximum .03782956/.72790479. The minimum observed forward slack is .00001136. These values describe the empirical Q; the wide bound range shows variable tightness. Checkpoints from one trajectory are correlated. The seeds, grid, and runs were fixed before outcome inspection.