BEYOND KERNEL DYNAMICS: LABEL INFLUENCE AT FEATURE EMERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear influence estimates the effect of training-label perturbations using the derivative of a learned predictor. We show that feature emergence can sharply restrict its range of accuracy, even when this derivative is computed exactly. We study factorized models and smooth, bias-free two-layer neural networks of fixed finite width trained from small initialization by squared-loss gradient flow and fixed-step gradient descent. Near first emergence, the label Jacobian separates into a rank-one timing term proportional to $L = \log(1/\epsilon)$ and a bounded remainder. Under nondegeneracy and, for gradient descent, stability and moment conditions, relative root-mean-square linearization error under isotropic Gaussian label noise vanishes if and only if $\sigma L \to 0$. At the critical scale, the response converges to a nonlinear random shift along the nominal trajectory. Propagated Gauss–Newton sensitivity misses the growing timing term; retaining curvature along the training direction restores it under gradient flow. We construct a derivative-preserving correction from the nominal trajectory and exact label response, prove critical-scale accuracy for exact trajectory evaluation, and give a finite-perturbation error bound for gradient flow. In a prospective test on width-64 networks, predictions from $L = 20$ transfer to $L = 40$ across three training configurations. A rate-based correction reduces retraining-forecast error by 70–83% at the primary emergence checkpoint and two prespecified noise levels, compared with exact linear influence. Multi-seed comparisons distinguish curvature-approximation error from finite-perturbation error and identify phase-dependent and small-noise failures.

1 INTRODUCTION

Suppose a network has reached a checkpoint, and we ask how its predictions would change if some training labels were different. Differentiating the training procedure gives an exact local answer. It does not establish how large a change that answer can describe. This distinction matters for sensitivity analysis and training-data attribution, where the quantity of interest is often a finite counterfactual rather than an infinitesimal derivative (Koh & Liang, 2017; Giordano et al., 2019; Bae et al., 2024).

We study this distinction near the first learning transition of a network initialized close to zero. The network initially remains near an unstable stationary point, then learns a feature over a much shorter interval. Labels determine the instability's growth rate. Consequently, a small label change can move the transition substantially while changing its shape only slightly. Evaluating the two runs at the same checkpoint then compares different phases of learning.

Our main result makes this timing mechanism precise for fully trained, smooth two-layer networks of any fixed finite width. If $L = \log(1/\epsilon)$, the prediction response has the expansion

$$D_y F_\epsilon = L\, v q^\top + K + o(1), \tag{1}$$

where v is the derivative of the emergence profile with respect to time, q is its clock sensitivity, and K remains bounded. The leading response is therefore rank one, even though every parameter is trained. At a fixed checkpoint, noise of size $1/L$ produces an order-one displacement along a nonlinear trajectory. This establishes a shrinking validity range for the *exact* derivative, independently of any approximation used to estimate it.

Influence peaks at learning transitions have been identified previously (Lee et al., 2026). Small-initialization analyses also describe stagewise nominal trajectories (Saxe et al., 2014; Pesme &

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.

Flammarion, 2023). Our contribution is the perturbation law through the transition: a uniform label-dependent entrance limit, its sharp noise boundary, and the separation between growing timing sensitivity and bounded content. We prove the entrance law for gradient flow and fixed-step gradient descent (GD), identify a broad class of curvature propagators that miss the timing term, and derive forecasts that preserve the exact derivative while representing a finite change in phase.

The empirical test follows these implications. After using initialization scales $L = 5, 10, 20$ to study the decomposition, we fixed quantitative predictions before running $L = 40$. This extension tests transfer in initialization scale, the phase window of timing dominance, and two nonlinear forecasts. It includes negative results: a velocity-projected clock becomes unreliable after the transition, and a stringent residual-norm prediction is slightly violated in two configurations. These observations delimit the mechanism rather than supporting an unrestricted claim about influence in deep networks.

2 THE COUNTERFACTUAL AND ITS DERIVATIVE

Let $X \in \mathbb{R}^{n\times d}$ contain the training inputs and $y \in \mathbb{R}^n$ the labels. We train $F(\theta) = (f_\theta(x_i))_{i=1}^n$ with squared loss,

$$\mathcal{L}(\theta; y) = \frac{1}{2n}\|F(\theta) - y\|^2, \qquad \dot\theta = -\nabla\mathcal{L}(\theta; y) \quad \text{or} \quad \theta_{k+1} = \theta_k - h\nabla\mathcal{L}(\theta_k; y). \tag{2}$$

Initialization, inputs, and the training schedule are shared between nominal and perturbed runs. Write $F_\epsilon(t; y)$ for training predictions and $A_\epsilon = D_y F_\epsilon(t_\epsilon; y_0)$ at a checkpoint selected from the nominal run. That checkpoint is held fixed under perturbation. This is the counterfactual of changing labels at a fixed training budget, including a stored checkpoint chosen retrospectively on the nominal trajectory. Stopping each perturbed run at its own progress level is a different estimand. Throughout, influence means a derivative with respect to labels, not deletion of examples or reweighting of their losses.

For $Z \sim N(0, I_n)$, define the actual change $\Delta_\epsilon = F_\epsilon(t_\epsilon; y_0 + \sigma Z) - F_\epsilon(t_\epsilon; y_0)$ and relative root-mean-square error

$$\mathcal{E}_\epsilon = \left(\frac{\mathbb{E}\|\Delta_\epsilon - \sigma A_\epsilon Z\|^2}{\mathbb{E}\|\Delta_\epsilon\|^2}\right)^{1/2}. \tag{3}$$

We study joint limits $\epsilon \to 0$ and $\sigma \to 0$, with data, width, and initialization direction fixed. The same derivatives can be evaluated on fresh inputs; their nondegeneracy and moment conditions are stated separately in the appendix.

Exact and curvature-approximated responses. Let $J_t = D_\theta F(\theta_t)$ and $R_t = \sum_i (F_i - y_i)\nabla^2_\theta F_i$. The parameter-label derivative $S_t = D_y\theta_t$ satisfies

$$\dot S_t = -\frac{1}{n}(J_t^\top J_t + R_t)S_t + \frac{1}{n}J_t^\top, \qquad S_0 = 0, \qquad A_t = J_t S_t. \tag{4}$$

Propagated Gauss–Newton (GN) replaces $J_t^\top J_t + R_t$ by $J_t^\top J_t$, while keeping the *same nominal trajectory and label forcing*. This distinguishes the approximation from training the network with a GN optimizer. The full initialization tangent instead freezes all parameter features at J_0. Our experiments compute these controls and the exact derivative separately.

3 A LABEL-DEPENDENT EMERGENCE CLOCK

Consider the bias-free network

$$f_\theta(x) = \sum_{j=1}^m a_j \phi(w_j^\top x), \qquad \theta_\epsilon(0) = \epsilon\xi, \tag{5}$$

where m is fixed, $\phi \in C^4$, $\phi(0) = 0$, and $s = \phi'(0) > 0$. Both layers are trained. Define

$$g(y) = X^\top y/n, \quad \lambda(y) = s\|g(y)\|, \quad L = \log(1/\epsilon), \quad T_L(y) = L/\lambda(y), \quad q = \frac{\nabla\lambda(y_0)}{\lambda_0^2}, \tag{6}$$

with $\lambda_0 = \lambda(y_0) > 0$. The linearized dynamics of each unit at the origin have eigenvalues $\lambda(y), 0, -\lambda(y)$. We assume the initialization direction has a nonzero projection onto the leading unstable eigenspace. The multiplicity of that eigenvalue is m, but its label-dependent rate is common to all units.

Theorem 1 (First-emergence response)**.** *Under the preceding assumptions, there is a nominal nonlinear emergence profile Φ such that, on compact sets of (τ, z),*

$$F_\epsilon(T_L(y_0) + \tau; y_0 + z/L) \longrightarrow \Phi(\tau + q^\top z) \quad \text{in } C^1. \tag{7}$$

At a fixed phase τ, the training-output response admits

$$A_\epsilon = L\Phi'(\tau)q^\top + K(\tau) + o(1). \tag{8}$$

Here K does not depend on L; the finite-L remainder is controlled in Appendix A. For Gaussian label noise, $\mathcal{E}_\epsilon \to 0$ if and only if $\sigma L \to 0$. At $\sigma L \to c \in (0, \infty)$,

$$F_\epsilon(T_L(y_0) + \tau; y_0 + \sigma Z) \longrightarrow \Phi(\tau + cq^\top Z) \quad \text{in } L^2, \tag{9}$$

and exact linear influence has a strictly positive limiting relative error.

For fixed-step GD, let $r_h(y) = \log(1 + h\lambda(y))$ and $q_h = \nabla r_h(y_0)/r_h(y_0)^2$. A real-phase profile Φ_h gives the corresponding compact-phase limits with $T_L = L/r_h$, $q = q_h$, and phase $\vartheta = k - L/r_h(y_0)$. In particular,

$$A_\epsilon(k) = L\Phi_h'(\vartheta)q_h^\top + K_h(\vartheta) + o(1). \tag{10}$$

The stochastic limit and sharp noise boundary additionally require nonzero profile velocity, post-emergence stability on a label neighborhood, and the moment envelope in Appendix C.

Proof mechanism. The incoming trajectory grows as $\epsilon e^{\lambda(y)t}$, reaching order one after $L/\lambda(y)$. Normalize time by $\lambda(y)$ before constructing and differentiating its nonlinear continuation. This yields a label-smooth profile with a bounded derivative at fixed phase. Differentiating the time argument contributes Lq and differentiating the profile contributes K. Uniform convergence of this construction through a label neighborhood of size $1/L$ gives (7). The bounded, nonaffine training-output profile then gives (9) and the sharp linearization boundary. In GD the incoming multiplier is $1 + h\lambda(y)$; a real-phase parametrization accounts for integer checkpoints without replacing GD by gradient flow. Appendices A–C give the details.

Scope of the GD extension. The local entrance law uses the discrete unstable multiplier and does not require a globally stable training trajectory. The Gaussian statement asks more: predictions must remain uniformly bounded for labels in a fixed neighborhood through the observation index $k_\epsilon = O(L)$, with a global envelope $\|F_\epsilon(k; y)\| \le C(1 + k)e^{c_E k}(1 + \|y\|)$. These conditions control perturbed runs that leave the transition early and the Gaussian tail outside the local neighborhood. Bounded activations supply the envelope, while a sufficiently small step on the relevant post-emergence region supplies local stability. Stability near the origin alone is insufficient. The theorem therefore separates a deterministic entrance calculation from the additional hypotheses needed for mean-square forecasts.

An explicit instance. For the scalar product $p = ab$, initial state $(a, b) = (\sqrt{\delta}, 0)$, and loss $(p - y)^2/2$, evaluate at the nominal midpoint $p(t_\delta; 1) = 1/2$. With $L_\delta = \log(4/\delta)$,

$$\partial_y p(t_\delta; 1) = \tfrac{1}{4}(L_\delta + 1) + o(1), \qquad p(t_\delta; 1 + cZ/L_\delta) \longrightarrow (1 + e^{-cZ})^{-1}. \tag{11}$$

Thus the nonlinear critical law is logistic-normal. At $\delta = 10^{-2}$, $1/L_\delta \simeq 0.17$; the exact finite-δ solution, rather than its asymptotic constant alone, is needed at this scale (Appendix G).

4 What the timing term implies

Geometry and attribution. Whenever the output velocity is nonzero, (8) implies

$$\|A_\epsilon\|_{\mathrm{op}} = \Theta(L), \qquad \mathrm{sr}(A_\epsilon) - 1 = O(L^{-2}), \qquad \eta_\epsilon := \frac{\|(I - \Pi_v)A_\epsilon\|_F^2}{\|A_\epsilon\|_F^2} = O(L^{-2}), \tag{12}$$

where $\mathrm{sr}(A) = \|A\|_F^2/\|A\|_{\mathrm{op}}^2$ and $\Pi_v = vv^\top/\|v\|^2$. The sensitivity peak approaches the speed peak when the latter is unique on the phase interval. Small η is evidence of alignment with the velocity; it is not a certificate of accuracy for finite perturbations.

The input factor is explicit: $q = XX^\top y_0/(n^2 s\|g(y_0)\|^3)$. For test inputs with nonzero limiting velocity, label-influence magnitudes therefore eventually rank examples by $|(XX^\top y_0)_i|$, assuming distinct scores. Their leading ranking is shared even when their predictions differ. Test-specific effects remain in K; test inputs with zero velocity are excluded from this statement.

A curvature approximation can miss an exact local effect. For a symmetric field H_t along the nominal path, define $\dot{S}^H = -H_t S^H/n + J_t^\top/n$. If $H_t \succeq 0$, contraction of its homogeneous propagator gives

$$\|S^H(t_\epsilon)\|_{\mathrm{op}} \le \frac{1}{n}\int_0^{t_\epsilon} \|J_t\|_{\mathrm{op}}\, dt = O(1). \tag{13}$$

The last bound uses the exponentially small incoming features and a fixed emergence phase. Thus $A^H(t_\epsilon) = J_{t_\epsilon} S^H(t_\epsilon) = O(1)$ while the exact response is $\Theta(L)$. In GD the same conclusion requires stable products of $I - hH_k/n$; $0 \preceq hH_k/n \preceq 2I$ is sufficient. Propagated GN satisfies the entrance bound under the stated stability conditions. This result concerns *sensitivity propagation on a fixed trajectory*; it does not compare training with different optimizers. Appendix D extends the bound to uniformly integrable negative curvature and examines a repair retaining curvature along the training direction.

Why updating the kernel is insufficient. The distinction is visible in the incoming unstable mode. Its linearized amplitude has the form $a(t; y) = \epsilon c(y) e^{\lambda(y)t}$. Differentiating at a fixed time gives

$$D_y a(t; y) = \epsilon e^{\lambda(y)t}\{D_y c(y) + t\, c(y) D_y \lambda(y)\}.$$

At emergence, the amplitude is order one but the waiting time is order L. The growing derivative comes from changing the growth rate throughout that waiting period. In the full variational equation, negative curvature amplifies the incoming label forcing at the same rate as the unstable mode, producing this extra factor of time. A contractive PSD propagator cannot accumulate that amplification, even when supplied with every nominal feature update. Thus the full initialization tangent and propagated GN omit different information: the former freezes the features, whereas the latter uses their evolution but omits the curvature responsible for the timing response. Corollary 2 identifies a sufficient correction to that curvature. The nonlinear forecast below addresses a separate issue: how to use the derivative once label changes are no longer infinitesimal.

Corollary 2 (Restoring the timing response)**.** *Under the gradient-flow assumptions of Theorem 1, let* $u_t = \dot{\theta}_t/\|\dot{\theta}_t\|$, $\Pi_t = u_t u_t^\top$, *and propagate sensitivity with*

$$H_t^{\mathrm{corr}} = J_t^\top J_t + \Pi_t R_t + R_t \Pi_t - \Pi_t R_t \Pi_t. \tag{14}$$

At a fixed emergence phase, $A_\epsilon^{\mathrm{corr}} - A_\epsilon = O(1)$. *Hence the corrected propagator recovers the leading order-*L *timing response.*

The added residual-curvature term has rank at most two. The corrected curvature H_t^{corr} agrees with the full Hessian along the parameter velocity and retains GN curvature on its orthogonal complement. Its action requires $R_t u_t$, not the complete residual Hessian; this vector can be obtained by a Hessian–vector product. The proof controls the rotation of the velocity direction during emergence (Appendix D.1). Recovering this local derivative does not remove the critical-scale failure of a linear forecast.

Retiming changes the estimand. If $\mathcal{T}_\epsilon(y)$ retimes each run, the chain rule gives

$$\widetilde{A}_\epsilon = A_\epsilon + v_\epsilon \nabla \mathcal{T}_\epsilon(y_0)^\top. \tag{15}$$

When $\mathcal{T}_\epsilon(y) = L/\lambda(y) + u_\epsilon(y)$ with bounded C^1 remainder, the order-L terms cancel. Matched-progress influence is bounded and has the same velocity-orthogonal component as fixed-horizon influence. With uniform C^2 control and a nonzero limiting content response, its relative linearization error vanishes for every $\sigma \to 0$, without requiring $\sigma L \to 0$ (Appendix E). This removes the timing restriction, not every source of nonlinearity.

What is retained after matching progress? Content is not a unique, clock-independent attribution. Two progress rules that agree at y_0, with bounded offsets $u_1(y)$ and $u_2(y)$, differ by $\widetilde{A}_1 - \widetilde{A}_2 = v_\epsilon \nabla(u_1 - u_2)^\top$. Only the velocity-orthogonal response is invariant to this choice. A training-loss crossing supplies a valid local progress rule when the loss decreases transversely through the chosen value; a nontransverse event requires separate analysis. Thus matching losses or peaks should be declared as part of the counterfactual, rather than introduced afterward to make a fixed-budget forecast look accurate. The fixed-horizon derivative measures what changes at the specified checkpoint. The matched-progress derivative answers what remains after allowing training time to change. Both questions are legitimate, but their leading attribution rankings can differ.

5 Forecasting a finite label change

The limit suggests using a nonlinear trajectory displacement for timing while preserving the exact local derivative. Let $H_0(t)$ interpolate the nominal predictions, $A = D_y F_\epsilon(t_0; y_0)$, and $w = H_0'(t_0)$. For a differentiable proposed shift $d(\delta y)$ with $d(0) = 0$, define a forecast of the prediction *change*:

$$P_d(\delta y) = H_0(t_0 + d(\delta y)) - H_0(t_0) + \{A - w\nabla d(0)^\top\}\delta y. \tag{16}$$

Its derivative at zero is exactly A. This identity requires using the interpolant's actual derivative, with consistent time units, output set, and anchor.

We compare two shifts. The *projected clock* estimates timing from the Jacobian,

$$d_{\mathrm{proj}}(\delta y) = \beta^\top \delta y, \qquad \beta = A^\top w/\|w\|^2. \tag{17}$$

The *rate clock* uses the analytic entrance rate,

$$d_{\mathrm{rate}}(\delta y) = L\left\{\frac{1}{r(y_0)} - \frac{1}{r(y_0 + \delta y)}\right\}, \quad r = \lambda \text{ (flow)}, \quad r = \log(1 + h\lambda) \text{ (GD)}. \tag{18}$$

Its gradient is Lq or Lq_h, so the linear residual in (16) subtracts precisely the proposed timing derivative. Both forecasts recover the critical law for exact nominal flow (Appendix F). For GD, (35) gives the corresponding limit for bounded rescaled perturbations when trajectory evaluation converges in C^1 to the real-phase profile. Gaussian RMS convergence additionally requires the stability and moment conditions in Appendix C.2 and uniform integrability of the squared forecast errors. A fixed-step Hermite interpolant preserves the derivative identity but can retain finite-shift interpolation error; we audit it separately. The projected clock divides by speed and can become unstable on a plateau; the rate clock does not. Neither construction guarantees an improvement when the perturbation is already small enough for linear influence to be accurate.

Proposition 3 (Finite-noise error of the rate forecast)**.** *Under the gradient-flow assumptions of Theorem 1, write the exact own-clock map as $W_\epsilon(y, s) = F_\epsilon(L/\lambda(y) + s; y)$ and let $t_0 = L/\lambda(y_0) + \tau$. Suppose the label segment from y_0 to $y_0 + \delta y$ and the phase segment from τ to $\tau + d$, where $d = d_{\mathrm{rate}}(\delta y)$, lie in a fixed compact product domain with $\lambda > 0$ and admissible times. Let M_{yy} and $M_{y\tau}$ bound $\|D_y^2 W_\epsilon\|$ and $\|\partial_\tau D_y W_\epsilon\|$ on that domain. Using the exact nominal trajectory in* (16)*,*

$$\|\Delta_\epsilon - P_{d_{\mathrm{rate}}}(\delta y)\| \le \tfrac{1}{2} M_{yy}\|\delta y\|^2 + M_{y\tau}|d|\,\|\delta y\|, \tag{19}$$

where $\Delta_\epsilon = F_\epsilon(t_0; y_0 + \delta y) - F_\epsilon(t_0; y_0)$.

To see the distinction from an ordinary Taylor remainder, use the exact identity $F_\epsilon(t_0; y_0 + \delta y) = W_\epsilon(y_0 + \delta y, \tau + d)$. Expand only the label argument at the shifted phase. The forecast retains that phase exactly and uses $K_\epsilon = D_y W_\epsilon(y_0, \tau) = A_\epsilon - L v_\epsilon q^\top$ for the remaining label effect. The second term in (19) accounts for the change of this content derivative between the two phases. Appendix F gives the calculation.

The own-clock derivatives are uniformly bounded on such domains. For a bounded rescaled perturbation $\delta y = z/L$, the shift is $O(1)$, and the bound is $O(1/L)$ even when the actual prediction change is order one. There is no additive finite-initialization error in this deterministic inequality. Its constants, however, concern nearby labels and are not an error certificate available from the nominal trajectory alone. For sampled GD, interpolation contributes an additional error; matching the derivative at the anchor does not remove it.

Evaluation at a checkpoint. For a specified counterfactual δy, the rate forecast can be evaluated in three steps.

1. Fix the output set and checkpoint, and compute the directional label response $a = A\delta y$ through the training updates. Retain the nominal outputs needed around that checkpoint. A full Jacobian is unnecessary for this single query; our experiments compute it to measure the complete response geometry.
2. Evaluate $r(y_0)$ and $r(y_0 + \delta y)$ to obtain the shift (18). Compute its derivative contribution $Lq^\top \delta y$ in flow time, or $Lq_h^\top \delta y$ in update units. The analytic rate is specific to the model assumptions; it is not inferred from the held-out perturbed outputs.
3. Query the nominal trajectory at the shifted time and add the residual directional response from (16). Use the same interpolation derivative in that residual and in the local derivative audit. Check trajectory coverage before reporting a forecast.

This procedure separates three potential errors: an inaccurate training derivative, an inaccurate clock, and an inaccurate shifted trajectory evaluation. The first is tested by label finite differences, the second by paired retraining, and the third by interpolation and coverage checks. The L40 experiment keeps these checks distinct. It also retains exact linear influence as a comparator at every noise level: derivative preservation establishes local agreement, not a guarantee that adding a nonlinear shift improves every finite perturbation.

These procedures require the nominal trajectory and label derivatives. They are controlled tests of a mechanism, not scalable replacements for all attribution methods. Nominal continuation may be needed to cover a positive shift; our tests bound this continuation and mark uncovered predictions unavailable. Neither forecast is fitted to the perturbed prediction being evaluated. A held-out shrinkage baseline tests whether their gains reduce to attenuating the linear forecast.

6 Sequential transitions in a diagonal network

The preceding MLP theorem addresses first emergence. A coupled diagonal model permits an extension to later transitions without assuming nonnegative feature correlations. Let $F = Xa$ with

$$\dot{a} = 2a \odot (b - Ga), \qquad a(0) = \epsilon \mathbf{1}, \qquad b = X^\top y/n, \quad G = X^\top X/n \succ 0. \tag{20}$$

Assume a strictly separated sequence of single-feature activations, positive least-squares coefficients on each active set, and no feature exits. For active columns X_S, let P_S be their orthogonal projector, and let x_j denote the next feature column. Define $c_j = (X_S^\top X_S)^{-1} X_S^\top x_j$ and $\rho_j = 2x_j^\top (I - P_S)y/n$. Then its emergence time is

$$T_j(y) = L\tau_j(y) + O(1), \qquad \tau_j = \frac{1 - \mathbf{1}^\top c_j}{\rho_j}, \qquad q_j = -\nabla \tau_j = \frac{\tau_j (I - P_S) x_j}{x_j^\top (I - P_S) y}. \tag{21}$$

Appendix H proves a rank-one timing response and the same critical scale at each such transition. The label direction now depends on the part of the emerging feature unexplained by the active features. The saddle-chain assumptions are substantive: positive-definite G alone does not guarantee this activation pattern. This result does not cover arbitrary later transitions of an MLP.

7 Predicting a new initialization scale

We test whether the timing description makes quantitative predictions beyond the runs used to construct it, and whether those predictions improve finite-perturbation forecasts. The main experiment is a prospective extension from $L \in \{5, 10, 20\}$ to $L = 40$. The scale, checkpoints, perturbations, methods, and evaluation rules were fixed before generating the extension. All comparisons use the same nominal parameters and perturbed labels.

Design. We train width-64, two-layer $\tanh$ networks by full-batch GD with step size $h = 0.05$. B1 has no biases and $n = 32$; B2 trains both biases with $n = 32$; B3 trains both biases with $n = 128$. The first 32 observations are shared, and all branches use 256 fixed fresh inputs. Non-bias

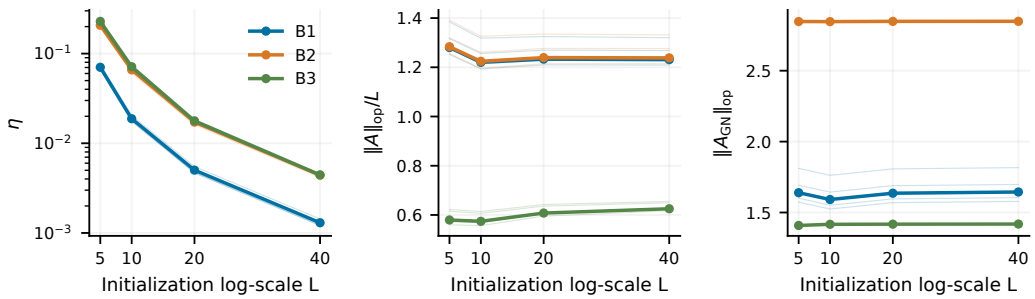


Figure 1: **Scale dependence at the primary checkpoint.** Measurements at $L = 5, 10, 20$ and the prospective test at $L = 40$. Thin curves show paired seeds; thick curves are branch medians. The exact response divided by L stays of comparable size while the unscaled GN response stays bounded on this range. These finite-scale observations do not establish asymptotic rates by themselves.

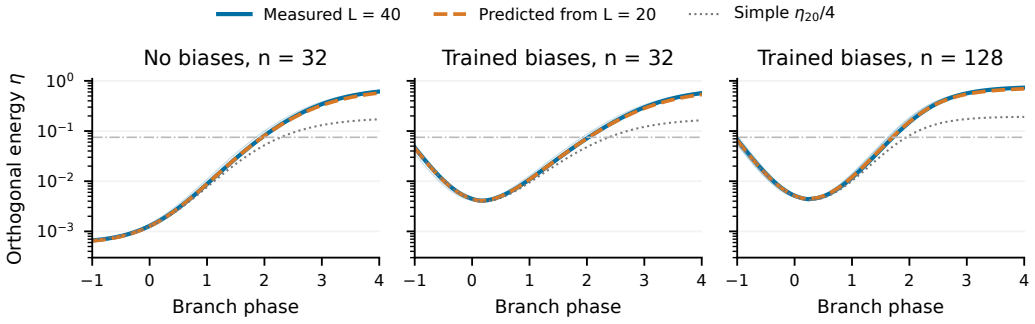


Figure 2: **A prospective geometry prediction.** The $L = 40$ orthogonal-energy curves are predicted using only $L = 20$ quantities and the analytic branch clock. Solid and dashed curves are seed medians; faint blue curves show all five paired seeds. The dotted baseline simply divides η_{20} by four. The horizontal line is the fixed window threshold 0.075. Biased branches use a centered analytic clock, without changing their training labels.

parameters start at $e^{-L}\xi$, with five paired Gaussian initialization seeds; biases start at zero. The response matrices include derivatives through every trained parameter. This is one synthetic data realization, with five initialization replicates, not a study across independently sampled datasets. B1 tests the theorem's architecture; B2 and B3 test extensions outside its bias-free assumptions.

For each nominal run we retrain on 16 fixed Gaussian label perturbations at four critical noise levels and one subcritical level, giving 15 nominal runs and 1,200 perturbed trajectories at $L = 40$. Requested phase u selects $k = \mathrm{round}((L + u)/r_{\mathrm{ref}})$, where $r_{\mathrm{ref}} = \log(1 + h\|X^\top y/n\|)$. Perturbed trajectories use these nominal integer checkpoints, never their own emergence times. Geometry plots instead use each branch's analytic phase, which differs slightly from this reference phase in B3. The measured η uses the update-slope proxy $w_k = J_k^{\mathrm{eval}}[-h\nabla\ell(\theta_k; y)]$; we do not assume that this equals the theorem's real-phase GD velocity. We report the relative RMS error in (3), aggregated over perturbations within a seed. The two primary comparison levels are $c = 0.3$ and 1, where $\sigma = c/(L\lambda_{32}\|q_{32}\|)$ and q_{32} is the continuous-time clock gradient. Appendix I gives the complete design and remaining noise levels.

The scale-dependent response. The scales $L = 5, 10, 20, 40$ show the predicted separation over the measured range (Figure 1). In B1, median $\|A\|_{\mathrm{op}}/L$ changes from 1.28 at $L = 5$ to 1.23 at $L = 40$, while the unscaled GN norm stays between 1.59 and 1.65 and median η falls from 0.0704 to 0.00130. These observations test finite-scale consequences of the decomposition; they do not prove its asymptotic rate or extend the theorem to biased networks.

Table 1: **Forecast errors at emergence.** Median relative RMS error over five paired initialization seeds at $L = 40$, requested phase 0, and $c = 1$. Lower is better. Fresh-output errors are measured against actual label-perturbed retraining.

Configuration	Linear	GN	Shrinkage	Projected	Rate
B1: no bias, $n = 32$	0.374	0.959	0.311	0.107	0.088
B2: bias, $n = 32$	0.430	0.960	0.361	0.109	0.088
B3: bias, $n = 128$	0.128	0.961	0.123	0.025	0.025

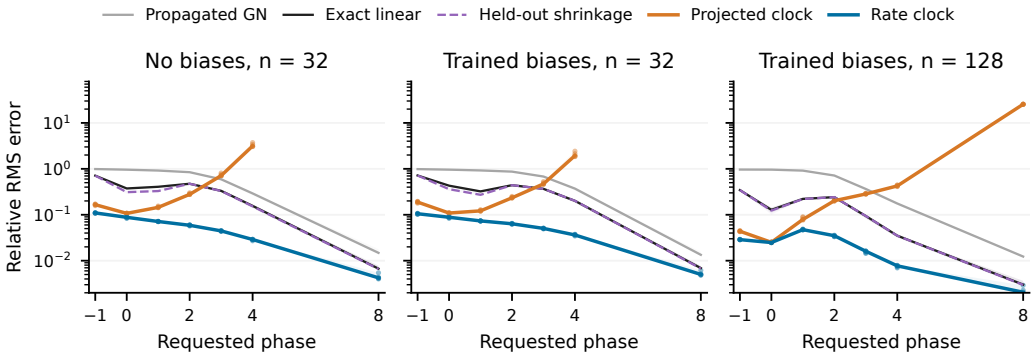


Figure 3: **Correction quality depends on phase and clock choice.** Fresh-output error at $L = 40$, $c = 1$. Lines are five-seed medians; pale bands span the five observed values. The rate correction remains accurate after the projected correction deteriorates. Projected forecasts at phase 8 are unavailable in B1/B2 because they exceed the fixed continuation cap; the curves stop rather than extrapolate. The $c = 0.3$ comparison is in Appendix I.

Geometry transfers to the unseen scale. We extrapolate the timing contribution while holding source-run residual norms fixed. Writing $C = \|A_{\perp,20}\|_F$ and $D = \|\Pi_{v_{20}} A_{20}\|_F - 20\|v_{20}\|\|q_h\|$, the forecast is

$$\widehat{\eta}_{40}(u) = \frac{C(u)^2}{C(u)^2 + [D(u) + 40\|v_{20}(u)\|\|q_h\|]^2}. \tag{22}$$

This scalar-norm extrapolation is motivated by the decomposition; it is not an identity implied by it. In particular, it uses no $L = 40$ velocity or fitted target-scale parameter. Over the prespecified phase grid $[-1, 4]$, log-scale R^2 is 0.9998, 0.9997, and 0.9998 in B1–B3. The simpler $\eta_{20}/4$ baseline gives 0.934, 0.877, and 0.846. Predicted window boundaries 2.00, 2.05, and 1.75 match at the declared 0.05 grid resolution. The stronger claim that every residual norm transfers within 1% is not fully supported: maximum deviations are 1.253%, 1.025%, and 0.362%. These are norm comparisons, not measurements of equality of the full content matrices.

Finite-noise forecasts. We compare exact linear influence, propagated GN, the full initialization tangent, zero change, the projected correction (17), and the rate correction (18). A shrinkage baseline fits a scalar multiplier of exact influence using only the other four initialization seeds, separately for each condition. Table 1 shows error magnitudes at the primary checkpoint. Across the 30 branch–seed–noise comparisons at this checkpoint, individual paired error reductions from the rate correction against exact linear influence range from 70% to 83%. GN remains near unit relative error despite using the evolving features.

At both primary noise levels, the projected correction beats linear influence, GN, and shrinkage in every seed through requested phase 2. At phases 3 and 4 it loses to linear influence and shrinkage. At phase 8 the same losses occur in all available comparisons; ten of the thirty comparisons are unavailable. In contrast, the rate correction beats all three baselines at every tested phase and both primary noise levels. These are repeated, paired comparisons within five seeds. The late advantage is small in absolute terms: at phase 8, the median paired error reduction against linear influence is 0.00079.

Table 2: **All noise levels, all seven phases, fresh outputs.** Entries count wins against the named baseline over $3 \times 5 \times 7 = 105$ repeated comparisons per noise level. These are descriptive paired counts, not 105 independent replicates. The last column counts unavailable projected forecasts. The primary levels were $c = 0.3, 1$; the others were retained as additional tests.

c	Rate/linear	Rate/GN	Rate/shrinkage	Projected/linear	Unavailable
0.0075	13/105	105/105	13/105	13/105	0
0.1	80/105	105/105	80/105	50/105	0
0.3	105/105	105/105	105/105	60/105	0
1	105/105	105/105	105/105	60/95	10
3	105/105	105/105	105/105	58/90	15

This does not make the rate correction a universal replacement for the derivative. At the subcritical level $c = 0.3/L = 0.0075$, exact linear influence is already accurate, and the rate correction wins only 13 of 105 branch–seed–phase comparisons. Finite-step interpolation can hurt a correction whose derivative at zero is nevertheless exact. A separate diagnostic at $c = 10^{-4}, 10^{-3}, 10^{-2}$ shows decreasing error as perturbations shrink. The chain-rule, label finite-difference, and sensitivity-decomposition checks meet their prespecified tolerances. The experiments therefore support a specific use: preserve a nonlinear timing shift when it materially affects the forecast, while retaining the exact linear predictor as the small-noise baseline.

8 Relation to earlier work and scope

Classical influence concerns optimized predictors (Koh & Liang, 2017; Giordano et al., 2019). Empirical fragility (Basu et al., 2021), counterfactual interpretation (Bae et al., 2022), and trajectory dependence (Schioppa et al., 2023) are established concerns. TracIn, TRAK, and SOURCE use checkpoints or approximate training dynamics for attribution (Pruthi et al., 2020; Park et al., 2023; Bae et al., 2024). Lee et al. (2026) connect influence peaks to stagewise learning through Bayesian influence and by differentiating solvable finite-time deep-linear trajectories. We characterize the exact label derivative's finite-perturbation range in nonlinear finite-width networks, its critical noise scale, and a derivative-preserving forecast.

Small-initialization analyses describe deep linear transitions (Saxe et al., 2014; Jacot et al., 2021), sequential diagonal learning (Pesme & Flammarion, 2023; Berthier, 2023), and nonlinear feature emergence (Kunin et al., 2025; Zhang et al., 2026). Silent alignment (Atanasov et al., 2022) and early neuronal alignment (Boursier & Flammarion, 2025) describe how small initialization organizes features. Section 6 analyzes label dependence along restricted diagonal saddle chains.

Kernel and lazy-training theories describe initialization-dependent regimes of nominal learning (Jacot et al., 2018; Chizat et al., 2019; Woodworth et al., 2020). Here a linear-kernel score determines the leading label direction; response magnitude and the finite-noise law depend on nonlinear dynamics.

9 Discussion

The actionable distinction is between a response caused by moving along a learning trajectory and one caused by changing that trajectory. At first emergence, the former grows with the waiting time and can invalidate both a PSD-curvature approximation and an exact linear forecast, for different reasons. A trajectory-aware forecast addresses the second failure without changing the local derivative.

The evidence is deliberately controlled: squared-loss regression, small smooth networks, one data realization, and five paired initialization seeds. The biased GD configurations test an extrapolation beyond the main theorem. Extreme initialization scales clarify the asymptotic mechanism but do not establish its prevalence in ordinarily initialized, stochastic, or large-scale models. Extending the analysis to later MLP transitions and deriving inexpensive diagnostics that avoid a full label Jacobian are the practical problems left open by these results.

REPRODUCIBILITY STATEMENT

The appendices specify the assumptions and proofs, data-generation and training protocols, evaluation procedures, and numerical checks. Predictions for the $L = 40$ experiment were fixed using $L = 20$ measurements before the target-scale runs were generated. The empirical analysis distinguishes initialization seeds from label-perturbation directions, retains missing forecasts, and reports departures from the specified predictions.

AI USE STATEMENT

Generative AI tools were used as research and engineering assistants in this work. They provided substantial assistance with implementing and debugging experimental code and running computational experiments, and were also used interactively for brainstorming, developing and checking mathematical and proof ideas, reviewing background literature, and assisting with manuscript and figure preparation and reference checking. The authors conceived and directed the research, determined which approaches and arguments to pursue, and were responsible for the final formulation and validation of the mathematical arguments and interpretation of the experimental results. AI-assisted outputs were critically reviewed and revised as needed by the authors, who take full responsibility for the final submission.

REFERENCES

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7234e0c36fdbcb23e7bd56b68838999b-Abstract-Conference.html.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training data attribution via approximate unrolling. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7af60ccb99c7a434a0d9d9c1fb00ca94-Abstract-Conference.html.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf/5f4c0d6ade226a87db8c80975d4487e28ec11d9e.pdf.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023. URL https://jmlr.org/papers/v24/22-1395.html.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

Xavier Cabré, Ernest Fontich, and Rafael de la Llave. The parameterization method for invariant manifolds I: Manifolds associated to non-resonant subspaces. *Indiana University Mathematics Journal*, 52(2):283–328, 2003a. URL https://web.mat.upc.edu/xavier.cabre/docs/indi1jour.pdf.

Xavier Cabré, Ernest Fontich, and Rafael de la Llave. The parameterization method for invariant manifolds II: Regularity with respect to parameters. *Indiana University Mathematics Journal*, 52(2):329–360, 2003b. URL https://web.mat.upc.edu/xavier.cabre/docs/indi2jour.pdf.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://dl.acm.org/doi/abs/10.5555/3454287.3454551.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A Swiss Army infinitesimal jackknife. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1139–1147, 2019. URL https://proceedings.mlr.press/v89/giordano19a.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2021. URL https://arxiv.org/abs/2106.15933.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/06cbd2e81dfbd3bb4cb0abce95b32584-Abstract-Conference.html.

Jin Hwa Lee, Matthew Smith, Maxwell Adam, and Jesse Hoogland. Influence dynamics and stagewise data attribution. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1edb917edb6cec59bf3a0afc03150e-Abstract-Conference.html.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/17a9ab4190289f0e1504bbb98d1d111a-Abstract-Conference.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Andrea Schioppa, Katja Filippova, Ivan Titov, and Polina Zablotskaia. Theoretical and practical perspectives on what influence functions do. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/57bb27b9be6ad04019ae3cea2b540872-Abstract.html.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine*

Learning Research, pp. 3635–3673, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Yedi Zhang, Andrew Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vit5M0G5Gb.

A PROOFS FOR THE NEURAL ENTRANCE LAW

This appendix states the regularity and observation conventions used in the main text and proves the neural results. All vector norms are Euclidean. Matrix norms without a subscript are operator norms. The data matrix has rows $x_i^\top$; the training-output vector is $F(\theta) = (f_\theta(x_i))_{i=1}^n$. We use

$$f_\theta(x) = \sum_{j=1}^m a_j \phi(w_j^\top x), \qquad \mathcal{L}(\theta; y) = \frac{1}{2n}\|F(\theta) - y\|^2, \qquad \theta(0) = \epsilon\xi, \qquad L = \log(1/\epsilon).$$

The width, data, initialization direction, and activation do not depend on ϵ. Assume $\phi \in C^4(\mathbb{R})$, $\phi(0) = 0$, and $s = \phi'(0) > 0$. Define

$$g(y) = X^\top y/n, \quad \lambda(y) = s\|g(y)\|, \quad \lambda_0 = \lambda(y_0) > 0, \quad q = \frac{\nabla\lambda(y_0)}{\lambda_0^2}.$$

All local uniformity statements are on a sufficiently small compact neighborhood U of y_0, on which λ remains positive. The nondegenerate-initialization condition is specified below. An evaluation vector on a fixed finite set of fresh inputs can replace F in deterministic conclusions. Statements involving a global prediction bound or Gaussian moments are stated for training outputs unless another envelope is assumed.

A.1 THE SADDLE AND A UNIFORM ENTRANCE CONSTRUCTION

Related invariant-manifold constructions and their parameter regularity are developed by Cabré et al. (2003a;b). The argument below is self-contained and establishes the label-derivative estimates required for influence.

Write the parameter flow as $\dot\theta = V(\theta, y) = -\nabla_\theta \mathcal{L}(\theta; y)$. Its linearization at zero is block diagonal across units, with blocks

$$M_j(y) = s\begin{pmatrix} 0 & g(y)^\top \\ g(y) & 0 \end{pmatrix}. \tag{23}$$

Each block has eigenvalues $\lambda(y), -\lambda(y)$ and $d-1$ zeros. Let $P_+(y), P_-(y), P_0(y)$ denote the orthogonal spectral projectors of the full matrix. These are smooth on U, including when the leading eigenvalue has multiplicity m. In particular, with $e(y) = g(y)/\|g(y)\|$,

$$(P_+(y)\xi)_j = (\alpha_j(y), \alpha_j(y)e(y)), \qquad \alpha_j(y) = \tfrac{1}{2}(\xi_{a_j} + e(y)^\top \xi_{w_j}).$$

We require $P_+(y_0)\xi \neq 0$. This condition persists on a smaller U and holds almost surely for an initialization direction with a density.

Normalize time before differentiating labels. Set $C(y) = M(y)/\lambda(y)$ and write the normalized vector field as

$$\frac{dz}{dr} = C(y)z + R(z, y), \qquad R(z, y) = O(\|z\|^2).$$

The eigenvalues of C are exactly $1, 0, -1$. Thus, for $a \geq 0$,

$$\begin{aligned} E_y(a) := e^{aC(y)} &= e^a P_+(y) + P_0(y) + e^{-a}P_-(y), \\ \|D_y^j E_y(a)\| \leq Ce^a, \qquad & \|D_y^j[E_y(a)(I - P_+(y))]\| \leq C, \end{aligned} \tag{24}$$

for $j = 0, 1, 2$. The last bound differentiates the product, so cancellation of the leading projector is preserved.

The required mixed derivatives of R exist under the stated activation assumption. Indeed, $V(\theta, y) = J(\theta)^\top y/n - J(\theta)^\top F(\theta)/n$ is affine in labels. After subtracting its linear part, its remainder is a linear combination, with smooth label-dependent coefficients, of C^3 functions that vanish to second order at zero. Consequently, locally and uniformly on U,

$$\begin{aligned} \|D_y^j R(z, y)\| \leq C\|z\|^2, \qquad & \|D_z D_y^j R(z, y)\| \leq C\|z\|, \\ \|D_z^2 D_y^j R(z, y)\| + \|D_z^3 D_y^j R(z, y)\| \leq C, \qquad & 0 \leq j \leq 2. \end{aligned} \tag{25}$$

Label derivatives here differentiate coefficients, rather than requiring higher activation derivatives.

Lemma 4 (Entrance with two label derivatives)**.** *There is a unique incoming normalized orbit ψ_y satisfying $e^{-r}\psi_y(r) \to P_+(y)\xi$ as $r \to -\infty$. Its finite-initialization counterpart $z_{\epsilon,y}$, initialized at $z_{\epsilon,y}(-L) = \epsilon\xi$, converges to ψ_y locally in C^2 in (r, y). On every compact set of phases and labels, the error is $O(\epsilon^\alpha)$ for every fixed $0 < \alpha < 1$. The constants can depend on this compact set and on α.*

Proof. For a negative r_*, use the weighted norm $\|z\|_* = \sup_{r\le r_*} e^{-r}\|z(r)\|$. The integral equation

$$\psi_y(r) = e^r P_+(y)\xi + \int_{-\infty}^{r} E_y(r-a)R(\psi_y(a), y)\, da \tag{26}$$

maps a fixed ball of this space into itself and has Lipschitz constant at most Ce^{r_*}. This follows by bounding its nonlinear integrand by $Ce^{r-a}e^{2a}$. Choose r_* so that this constant is less than $1/2$. Equation (25) and the summable majorant $Ce^{r-a}e^{2a}$ show that the map and its first two label derivatives are continuous in the weighted norm. The second derivative of a composition is

$$D_z R\, D_y^2\psi + D_z^2 R[D_y\psi, D_y\psi] + 2D_z D_y R\, D_y\psi + D_y^2 R.$$

Once $\psi, D_y\psi = O(e^r)$, every forcing term after the first is $O(e^{2r})$. Inverting $I - D_\psi\mathcal{T}$ by its convergent Neumann series proves twice differentiable dependence and the bounds

$$D_y^j\psi_y(r) = e^r D_y^j(P_+(y)\xi) + O(e^{2r}), \qquad j = 0, 1, 2. \tag{27}$$

Time and mixed derivatives follow from the differential equation. Any solution with the specified backward limit satisfies (26): the complementary homogeneous modes cannot have a finite limit after division by e^r, and variation of constants removes them. Contraction then gives uniqueness.

We give the comparison argument because an undifferentiated trajectory limit would not justify influence. The finite path has mild equation

$$z_{\epsilon,y}(r) = E_y(r+L)\epsilon\xi + \int_{-L}^{r} E_y(r-a)R(z_{\epsilon,y}(a), y)\, da.$$

Equations (24)–(25), first for the state and then for its label derivatives, give $\|D_y^j z_{\epsilon,y}(r)\| \le Ce^r$ for $-L \le r \le r_*$ and $j \le 2$. To see this without a time-dependent Gronwall loss, divide by e^r; the only coefficient multiplying a weighted unknown is Ce^a, whose integral up to r_* is uniformly small. The remaining derivative terms are bounded forcings with the same integrable weight.

Let $e = z_{\epsilon,y} - \psi_y$ and $E_j(r) = e^{-r}\|D_y^j e(r)\|$. At $r = -L$, the difference is $\epsilon(I - P_+)\xi + O(\epsilon^2)$, with the same order after two label derivatives. Propagating this expression as a whole using (24) and differentiating the mild difference equation yields

$$E_j(r) \le Ce^{-(r+L)} + C\epsilon + C\int_{-L}^{r} e^a \sum_{i=0}^{j} E_i(a)\, da, \qquad j = 0, 1, 2. \tag{28}$$

For example, the second-composition differences are bounded by $Ce^a\sum_{i\le 2}\|D_y^i e(a)\|$ using bounded third state derivatives in (25). No nonintegrable coefficient is introduced by differentiation. Sum (28) over j and apply the integral resolvent form of Gronwall's inequality. At a fixed r_*, the contribution of the decaying forcing to the integral is bounded by $C\int_{-L}^{r_*} e^a e^{-(a+L)}da = C(L + r_*)\epsilon$. The resolvent is bounded because $\int_{-L}^{r_*} e^a da$ is bounded. Thus the comparison is $O((1 + L)\epsilon)$ at r_*, and in particular $O(\epsilon^\alpha)$ for every $\alpha < 1$. The differential equation controls time and mixed derivatives. Smooth finite-time continuation gives the claim on later compact intervals. □

A.2 From entrance to influence

Return to physical time by $\Psi_y(\tau) = \psi_y(\lambda(y)\tau)$. Set

$$W_\epsilon(y, \tau) = F_\epsilon(L/\lambda(y) + \tau; y), \qquad \Gamma(y, \tau) = F(\Psi_y(\tau)), \qquad \Phi(\tau) = \Gamma(y_0, \tau).$$

Lemma 4 proves $W_\epsilon \to \Gamma$ locally in C^2 at an exponential rate in L. The incoming orbit extends to every finite positive phase. In fact, any squared-loss gradient flow satisfies

$$\int_0^t \|\dot\theta\|^2\, ds \le \mathcal{L}(\theta(0); y), \qquad \|\theta(t) - \theta(0)\| \le \sqrt{t\mathcal{L}(\theta(0); y)},$$

which precludes finite-time parameter blowup. Loss descent also gives the global training-output envelope

$$\|F_\epsilon(t;y)\| \le 2\|y\| + \|F(\epsilon\xi)\|, \qquad \|\Phi(\tau)\| \le 2\|y_0\|. \tag{29}$$

The incoming orbit is not stationary. If its prediction velocity vanished at a finite phase, then $\Phi' = -JJ^\top(\Phi - y_0)/n = 0$ would imply $J^\top(\Phi - y_0) = 0$ by taking the inner product with the residual. The parameter flow would then be stationary, contradicting uniqueness and the nonzero incoming coefficient. Thus $\Phi'(\tau) \neq 0$ for training outputs at every finite phase.

At the fixed observation time $t_\epsilon = L/\lambda_0 + \tau$, the exact own-clock identity is

$$F_\epsilon(t_\epsilon; y_0 + u) = W_\epsilon(y_0 + u, \tau + d_L(u)), \quad d_L(u) = L\{\lambda_0^{-1} - \lambda(y_0 + u)^{-1}\}. \tag{30}$$

Since $d_L(v/L) = q^\top v + O(\|v\|^2/L)$ locally in C^1, this proves the main entrance law. Differentiating (30) at zero gives the exact finite-initialization decomposition

$$A_\epsilon = Lv_\epsilon q^\top + K_\epsilon, \qquad v_\epsilon = \partial_\tau W_\epsilon(y_0, \tau), \qquad K_\epsilon = D_y W_\epsilon(y_0, \tau). \tag{31}$$

Here $v_\epsilon \to \Phi'(\tau)$ and $K_\epsilon \to D_y\Gamma(y_0, \tau)$ exponentially in L. The coefficients at finite initialization may depend on L. Equivalently, $A_\epsilon = L\Phi'(\tau)q^\top + D_y\Gamma(y_0, \tau) + O(\epsilon^\alpha)$ after reducing $\alpha < 1$ to absorb polynomial factors. This is the precise meaning of the affine asymptotic expansion.

For completeness,

$$q_i = \frac{(XX^\top y_0)_i}{n^2 s\|g(y_0)\|^3}, \qquad q \neq 0. \tag{32}$$

Nonvanishing follows because $g \in \mathrm{range}(X^\top)$ and $Xg = 0$ would imply $g = 0$.

Weyl's singular-value inequality applied to (31) gives $\sigma_1(A_\epsilon) = L\|v_\epsilon\|\|q\| + O(1)$ and $\sigma_j(A_\epsilon) = O(1)$ for $j \ge 2$. Therefore

$$\frac{\|A_\epsilon\|_F^2}{\|A_\epsilon\|^2} - 1 = O(L^{-2}), \qquad \frac{\|(I - P_{v_\epsilon})A_\epsilon\|_F^2}{\|A_\epsilon\|_F^2} = O(L^{-2}), \quad P_v = vv^\top/\|v\|^2.$$

Uniform convergence on a compact phase interval gives uniform convergence of $\|A_\epsilon\|/L$ to $\|q\|\|\Phi'\|$. A unique maximizer of the latter is consequently the limit of any maximizer of the former on that interval. This is a statement about a declared compact window, not a guarantee that a finite training horizon contains a global peak.

B THE NOISE SCALE AND ITS QUALIFICATIONS

Let $Z \sim N(0, \Sigma)$ with $q^\top \Sigma q > 0$, let $\sigma \to 0$, and define

$$\Delta_\epsilon = F_\epsilon(t_\epsilon; y_0 + \sigma Z) - F_\epsilon(t_\epsilon; y_0), \qquad \mathcal{E}_\epsilon = \frac{\|\Delta_\epsilon - \sigma A_\epsilon Z\|_{L^2}}{\|\Delta_\epsilon\|_{L^2}}.$$

Write $\kappa = \sigma L$ and $H_\epsilon(v) = F_\epsilon(t_\epsilon; y_0 + v/L)$. The entrance result gives compact C^1 convergence to $H(v) = \Phi(\tau + q^\top v)$ and $A_\epsilon/L \to DH(0) = \Phi'(\tau)q^\top$.

If $\kappa \to 0$, on $\{\kappa\|Z\| \le 1\}$ the fundamental theorem of calculus yields

$$\frac{\Delta_\epsilon}{\kappa} = \int_0^1 DH_\epsilon(a\kappa Z)Z\, da \longrightarrow \Phi'(\tau)q^\top Z.$$

The integrand is dominated by $C\|Z\|$. On the complement, the envelope (29) bounds the squared normalized change by $C\kappa^{-2}(1 + \sigma^2\|Z\|^2)$. Gaussian tails beyond $1/\kappa$ make its expectation tend to zero. Thus the convergence holds in L^2. The normalized linear forecast has the same limit, whose norm is positive, proving $\mathcal{E}_\epsilon \to 0$ even when noise is smaller than the entrance approximation error.

If $\kappa \to c \in (0, \infty)$, the same envelope and pointwise convergence give

$$\Delta_\epsilon \longrightarrow D_c := \Phi(\tau + cq^\top Z) - \Phi(\tau), \qquad \mathcal{E}_\epsilon \longrightarrow \frac{\|D_c - c\Phi'(\tau)q^\top Z\|_{L^2}}{\|D_c\|_{L^2}} > 0. \tag{33}$$

The denominator is nonzero because $q^\top Z$ has full support on $\mathbb{R}$ and Φ is nonconstant. A zero numerator would make the bounded function Φ equal a nonconstant affine function on all of $\mathbb{R}$,

which is impossible. If $\kappa \to \infty$, the true change remains bounded in L^2, while the linear forecast has norm asymptotic to $\kappa\|\Phi'(\tau)q^\top Z\|_{L^2}$. The reverse triangle inequality gives $\mathcal{E}_\epsilon \to \infty$. Subsequence extraction proves the necessary and sufficient condition $\mathcal{E}_\epsilon \to 0$ if and only if $\sigma L \to 0$.

This argument is a singular version of the ordinary delta method, with a model-specific entrance map supplying the non-affine limit. A covariance supported in $\ker q^\top$ is deliberately excluded. In such directions the leading timing term vanishes, and a different scale can govern the first nonzero phase shift.

C Fixed-step gradient descent

Let $G_y(\theta) = \theta - h\nabla\mathcal{L}(\theta; y)$ for a fixed $h > 0$. Its linearization is $B_h(y) = I + hM(y)$, and its remainder R_h is $O(\|\theta\|^2)$. Put

$$\mu(y) = 1 + h\lambda(y), \quad r_h(y) = \log \mu(y), \quad \nu(y) = \max\{1, |1 - h\lambda(y)|\}, \quad q_h = \frac{\nabla r_h(y_0)}{r_h(y_0)^2}. \tag{34}$$

The strict inequality $\nu < \mu$ holds for every $h > 0$. Therefore a local entrance law does not require global GD stability. The latter is needed separately for Gaussian moments at checkpoints that can lie well after a perturbed trajectory's emergence.

C.1 A real-phase family and the discrete entrance estimate

There is a C^2 family $\Psi_{h,y}(\rho)$ with

$$\Psi_{h,y}(\rho + 1) = G_y(\Psi_{h,y}(\rho)), \qquad \mu(y)^{-\rho}\Psi_{h,y}(\rho) \to P_+(y)\xi \quad (\rho \to -\infty). \tag{35}$$

The real variable ρ indexes a family of integer-step orbits; it is not a change in the GD update rule. To construct it, solve

$$\Psi_{h,y}(\rho) = \mu(y)^\rho P_+(y)\xi + \sum_{j\geq 1} B_h(y)^{j-1} R_h(\Psi_{h,y}(\rho - j), y). \tag{36}$$

Choose $0 < r_1 < \min_U r_h \leq \max_U r_h < r_2 < 2r_1$. In the norm $\sup_{\rho\leq\rho_*} e^{-r_1\rho}\|z(\rho)\|$, the nonlinear series and its Lipschitz constant are bounded by a constant times $e^{r_1\rho_*}\sum_{j\geq 1} e^{r_2(j-1)-2r_1 j}$. This is a contraction for negative enough ρ_*. First and second label derivatives of B_h^{j-1} contribute at most a polynomial in j; those of μ^ρ contribute at most a polynomial in $|\rho|$. The strict exponential margins make all differentiated series summable. The mixed bounds (25) apply to R_h as well, proving twice differentiable dependence on labels and incoming amplitude.

Phase smoothness also follows from this amplitude dependence. If $\Psi_{y,p}$ denotes the construction with incoming amplitude $P_+(y)p$, uniqueness gives $\Psi_{y,p}(\rho + a) = \Psi_{y,\mu(y)^a P_+(y)p}(\rho)$ for $a \leq 0$. Evaluating at ρ_* expresses earlier phases as a C^2 function of the smooth amplitude $\mu^{\rho-\rho_*}P_+\xi$. Equation (35), obtained by separating the first summand, extends the family forward smoothly by finitely many GD steps. Uniqueness holds on each coset of $\mathbb{R}/\mathbb{Z}$, since any orbit with the same incoming coefficient satisfies the lattice restriction of (36).

The weighted contraction initially bounds the nonlinear remainder by $O(e^{2r_1\rho})$. Since $2r_1 > r_h(y)$, this first implies $\Psi_{h,y}(\rho) = O(\mu(y)^\rho)$. Substitution back into (36) then improves the remainder to $O(\mu(y)^{2\rho})$. Its first two label derivatives have the same exponential order with polynomial factors in $1 + |\rho|$. These factors will be included in the polynomial-in-L comparison estimates below.

Here are explicit comparison estimates. Let $T(y) = L/r_h(y)$, $\rho_k = k - T(y)$, and $e_k = \theta_\epsilon(k; y) - \Psi_{h,y}(\rho_k)$. Since $\mu^{-T} = \epsilon$ identically in labels,

$$e_0 = \epsilon(I - P_+)\xi + O(\epsilon^2), \qquad e_k = B_h^k e_0 + \sum_{i<k} B_h^{k-1-i}\{R_h(\theta_i, y) - R_h(\Psi_{h,y}(\rho_i), y)\}.$$

In the entrance region $\rho_k \leq \rho_*$, a bootstrap gives $\|\theta_k\| \leq C\mu^{\rho_k}$. With $U_k = \mu^{-\rho_k}\|e_k\|$, the last display implies

$$U_k \leq C(\nu/\mu)^k + C\epsilon + C\sum_{i<k} \mu^{\rho_i} U_i.$$

The sum of its coefficients is bounded, and retaining the decaying forcing in the discrete resolvent gives $U_k \le C(\nu/\mu)^k + C\epsilon(1 + k\nu^k)$. At an index with bounded phase, this is $O(L\epsilon^a)$, where $a = 1-\log\nu/\log\mu > 0$. Differentiation of the initial term uses $\|D_y^j(B_h^k(I-P_+))\| \le C(1+k)^j\nu^k$. For the nonlinear difference, the first derivative is bounded by $C\mu^{\rho_i}(\|D_y e_i\| + L\|e_i\|)$, because the own-clock phase has label derivative $O(L)$. A second differentiation has the same coefficient on $D_y^2 e_i$ and polynomial-in-L multiples of lower-order errors. Discrete resolvent iteration therefore gives, for $j \le 2$, $\|D_y^j e_k\| \le CL^{p_j}\epsilon^a$ on bounded phases, for fixed finite integers p_j. Finitely many remaining GD steps preserve these bounds. Thus, on smaller U and for every $\alpha < \inf_U a$, they are $O(\epsilon^\alpha)$ after reducing α if necessary.

Let $\vartheta = k - L/r_h(y_0)$ and $\Phi_h = F \circ \Psi_{h,y_0}$. Since $k - L/r_h(y_0 + v/L) = \vartheta + q_h^\top v + O(\|v\|^2/L)$, we obtain compact C^1 convergence

$$F_\epsilon(k; y_0 + v/L) \to \Phi_h(\vartheta + q_h^\top v), \qquad A_\epsilon(k) = L\Phi_h'(\vartheta)q_h^\top + K_h(\vartheta) + O(\epsilon^\alpha), \tag{37}$$

where $K_h(\vartheta) = D_y[F(\Psi_{h,y}(\vartheta))]_{y_0}$ and the error is uniform over bounded ϑ. Integer observation times generally give an oscillating phase; no phase limit is required for this uniform statement. Direct expansion gives

$$hq_h = q\frac{(h\lambda_0)^2}{(1 + h\lambda_0)\log^2(1 + h\lambda_0)} = q\{1 + (h\lambda_0)^2/12 + O(h^3)\}.$$

C.2 What is required for random labels under GD

For clarity we separate three statements. The deterministic compact entrance law (37) holds without a post-emergence stability condition. For Gaussian perturbations, assume there are $\delta, C > 0$ such that training predictions up to the observation index $k_\epsilon = O(L)$ are bounded by C for all $\|y-y_0\| \le \delta$. Also assume a global envelope

$$\|F_\epsilon(k; y)\| \le C(1 + k)e^{c_E k}(1 + \|y\|), \qquad k \le k_\epsilon. \tag{38}$$

On a compact phase window where $\|\Phi_h'\|$ is bounded away from zero, these assumptions imply subcritical consistency and the critical random-phase law for $\sigma L = O(1)$. Local stability also bounds the entire limiting profile: any fixed phase shift can be realized by a perturbation v/L with fixed v, so the compact entrance limit implies $\|\Phi_h(\rho)\| \le C$ for every real ρ. The non-affinity argument in Appendix B therefore applies. Inside the label ball the same moment proof goes through. Outside it, Gaussian tails give

$$\mathbb{E}[\|\Delta_\epsilon\|^2 \mathbf{1}_{\{\sigma\|Z\|>\delta\}}] \le CL^2 \exp(C_E L - c/\sigma^2).$$

This is negligible at subcritical and critical noise, including after division by $(\sigma L)^2$. If $\sigma L \to \infty$, it also proves divergence when additionally $\sigma^2 L \to 0$ if $C_E > 0$. Without that tail restriction, the necessary part of the iff still follows by a different argument. On the label ball, local stability gives $\Delta_\epsilon/(\sigma L) \to 0$ in probability. The probability of leaving the ball tends to zero, so this convergence holds globally. In contrast, $\sigma A_\epsilon Z/(\sigma L)$ converges along each convergent phase subsequence to a nonzero Gaussian vector. If the relative RMS error tended to zero, the reverse triangle inequality would imply

$$\frac{\|\Delta_\epsilon\|_{L^2}}{\|\sigma A_\epsilon Z\|_{L^2}} \to 1, \qquad \frac{\|\Delta_\epsilon - \sigma A_\epsilon Z\|_{L^2}}{\sigma L} \to 0.$$

This would force $\Delta_\epsilon/(\sigma L)$ to have the same nonzero probability limit as the linear forecast, a contradiction. Each moment is finite by (38). Together with the strictly positive critical limit, the subsequence argument proves the full iff under the stated local stability and envelope conditions. What remains conditional on the stronger tail control is supercritical divergence to infinity, rather than failure to converge to zero.

These hypotheses can be checked or imposed independently of the saddle calculation. For bounded $|\phi| \le c_\phi$, the readout update yields (38) with $e^{c_E} = \max\{1, hmc_\phi^2 - 1\}$, up to a constant and the factor $1 + k$. A sufficiently small local step along every segment between consecutive parameters gives loss descent on the label ball and hence local stability. Neither condition is automatic from $h\lambda_0 < 2$: the curvature after emergence can be much larger than at the origin.

D CURVATURE PROPAGATION AND THE MISSING TIMING MODE

Let $J_t = D_\theta F(\theta_t)$, $e_t = F(\theta_t) - y_0$, and $H_t^{\mathrm{ex}} = J_t^\top J_t + \sum_i e_{t,i} D_\theta^2 F_i(\theta_t) = n D_\theta^2 \mathcal{L}(\theta_t; y_0)$. For a symmetric curvature field H_t along the nominal trajectory, define

$$\dot{S}_t^H = -H_t S_t^H/n + J_t^\top/n, \quad S_0^H = 0, \qquad A_t^H = J_t S_t^H. \tag{39}$$

This is an approximation to the derivative of the specified gradient-flow training procedure. It does not define training by a different optimizer.

Let $N_t = \int_0^t \|(H_s)_-\|\, ds$, where $(H_s)_-$ is the negative part. The homogeneous propagator obeys $\|U(t,s)\| \le \exp((N_t - N_s)/n)$, obtained by differentiating $\|U(t,s)z\|^2$. Variation of constants therefore gives

$$\|S_t^H\| \le \frac{e^{N_t/n}}{n} \int_0^t \|J_s\|\, ds. \tag{40}$$

In the bias-free network $J(0) = 0$. Lemma 4 gives $\|\theta_t\| \le C e^{\lambda_0 t - L}$ until a fixed negative emergence phase, so $\int \|J_t\| dt = O(1)$ there. The remaining fixed-length interval has bounded Jacobians. Thus $\int_0^{L/\lambda_0+\tau} \|J_t\| dt = O(1)$ uniformly on compact phase intervals. If $N_t = O(1)$, equation (40) implies $\|A_t^H\| = O(1)$; in particular this holds for every positive-semidefinite H_t. At critical noise its forecast vanishes, whereas the true change has a nonzero random-phase limit. Its relative RMS prediction error tends to one and its predicted total variance tends to zero.

The same conclusion holds at fixed fresh inputs because their terminal parameter Jacobian is bounded. In contrast, the Gaussian relative-error conclusion there requires a suitable fresh-output envelope and a nonconstant limiting evaluation profile.

The exact layer-block-diagonal Hessian is also covered, although it need not be positive semidefinite. The readout-readout residual Hessian is zero, and the hidden-hidden residual Hessian has blocks $\sum_i e_i a_j \phi''(w_j^\top x_i) x_i x_i^\top$. Its negative part is at most $C\|a\|\|e\|$. The residual is bounded by loss descent and $\|a\|$ is integrable on the entrance interval, so $N_t = O(1)$. At the origin the entire nonzero loss Hessian lies in cross-layer blocks. Discarding those blocks removes the sustained negative curvature that transports the timing mode.

For fixed features $\widehat{J}$, the training-output response is $I - e^{-\widehat{J}\widehat{J}^\top t/n}$. Initialization features satisfy $\|\widehat{J}\| = O(\epsilon)$, so their response at $t = O(L)$ is $O(\epsilon^2 L)$.

Discrete propagators. For $s_{k+1} = (I - hH_k/n)s_k + (h/n)J_k^\top z$, the condition $0 \preceq hH_k/n \preceq 2I$ gives $\|s_k\| \le (h/n)\sum_{j<k} \|J_j\|\|z\|$. The entrance bound $\|J_j\| \le C\mu^{j - L/r_h}$ makes this sum $O(1)$ at emergence. The same result permits a fixed number of final factors with uniformly bounded norms. In particular, propagated GN has the bound for any fixed h: its factors are contractions sufficiently far inside the entrance region because $J_j \to 0$, and only a bounded number of subsequent steps lie in a fixed phase window. Arbitrary positive-semidefinite matrices without discrete stability need not satisfy it.

An estimate requiring no entrance structure is useful for comparison. If $H_t \succeq cJ_t^\top J_t$ and $u = S_t^H z$, then

$$\frac{d}{dt}\|u\|^2 \le -\frac{2c}{n}\|J_t u - z/(2c)\|^2 + \frac{\|z\|^2}{2cn}, \qquad \|S_t^H\| \le \sqrt{t/(2cn)}.$$

It follows by completing the square. This weaker $O(\sqrt{L})$ estimate remains applicable when a trained output bias makes $J(0) \neq 0$.

D.1 A SUFFICIENT REPAIR OF THE CURVATURE MODEL

There is a converse for continuous-time propagation. Let $w_t = \dot{\theta}_t$, $e_t^\theta = w_t/\|w_t\|$, and assume H_t' is symmetric and uniformly bounded, satisfies $H_t' w_t = H_t^{\mathrm{ex}} w_t$, and is positive semidefinite on $w_t^\perp$. Then

$$S_{t_\epsilon}^{H'} - S_{t_\epsilon} = O(1), \qquad A_{t_\epsilon}^{H'} = A_{t_\epsilon} + O(1). \tag{41}$$

We provide the argument, including the moving-direction term. Differentiating the normalized-time representation gives $S_t = t\lambda_0 w_t q^\top + E_t$, where $\|E_t\| \le C e^{\lambda_0 t - L}$ in the entrance region and $E_t =$

$O(1)$ in the remaining compact window. The difference $D = S - S^{H'}$ solves $\dot{D} = -H'D/n + f$, with $f = -(H^{\mathrm{ex}} - H')E/n$. Symmetry gives $(e_t^\theta)^\top f = 0$. Write $D = e_t^\theta b + D_\perp$. Transverse positivity gives $\|D_\perp(t)\| \le \int_0^t \|f_s\| ds$. Since $\dot{w} = -H^{\mathrm{ex}} w/n = -H'w/n$, direct differentiation of this moving decomposition gives

$$b(t) = 2\|w_t\| \int_0^t \frac{(\dot{e}_s^\theta)^\top D_\perp(s)}{\|w_s\|}\, ds.$$

On the entrance region the unstable projection supplies $\|w_t\| \ge c e^{\lambda_0 t - L}$, while the stable and quadratic remainders give $\|\dot{e}_t^\theta\| \le C(e^{-2\lambda_0 t} + e^{\lambda_0 t - L})$. The latter has bounded integral. Moreover $\|D_\perp(t)\| \le C e^{\lambda_0 t - L}$. On the remaining compact window w_t is bounded away from zero and all terms are bounded. The last display therefore bounds b, proving (41).

One admissible repair is

$$H_t^\flat = J_t^\top J_t + P_t R_t + R_t P_t - P_t R_t P_t, \quad P_t = e_t^\theta (e_t^\theta)^\top, \quad R_t = H_t^{\mathrm{ex}} - J_t^\top J_t.$$

It is exact on w_t and has transverse quadratic form $\|J_t x\|^2$. This is a structural sufficient condition for recovering the leading timing coefficient. It is not a claim that an implemented discrete optimizer or a single Hessian eigenpair achieves the same guarantee.

E Retiming, content, and attribution

Let a differentiable progress rule choose $T_\epsilon(y)$ with $T_\epsilon(y_0) = t_\epsilon$. The chain rule gives

$$\widetilde{A}_\epsilon := D_y[F_\epsilon(T_\epsilon(y); y)]_{y_0} = A_\epsilon + v_\epsilon \nabla T_\epsilon(y_0)^\top. \tag{42}$$

Consequently $(I - P_{v_\epsilon})\widetilde{A}_\epsilon = (I - P_{v_\epsilon})A_\epsilon$ for every progress rule. If $T_\epsilon(y) = L/\lambda(y) + u_\epsilon(y)$ with bounded first derivative of u_ϵ, then $\widetilde{A}_\epsilon = K_\epsilon + v_\epsilon \nabla u_\epsilon(y_0)^\top = O(1)$. Only the component orthogonal to velocity is invariant to the choice of progress functional.

Lemma 4 supplies the stronger regularity needed for a small-noise statement. If $u_\epsilon \to u$ locally in C^2 and the limiting matched-progress derivative is nonzero on the noise covariance, then $G_\epsilon(y) = W_\epsilon(y, u_\epsilon(y))$ has uniformly bounded second derivatives. Taylor's theorem bounds its RMS linearization residual by $C\sigma^2(\mathbb{E}\|Z\|^4)^{1/2}$ locally. The matched-progress change has RMS norm asymptotic to a positive constant times σ. Outside the label neighborhood, the all-time envelope (29) and Gaussian tails are negligible; choose any nonnegative measurable extension of the progress times there. Therefore the relative error is $O(\sigma)$ for every $\sigma \to 0$, without requiring $\sigma L \to 0$. This removes the timing-induced shrinking scale. It does not establish accuracy for fixed large noise or when the limiting content derivative vanishes.

Examples follow from the implicit function theorem applied in own-clock coordinates. For a target training-loss value attained at a fixed nominal phase, the phase derivative is strictly negative, so a nearby hitting time has the required form and converges in C^2. A prediction-norm threshold requires a transverse first crossing. A speed maximum requires a unique nondegenerate maximizer and enough smoothness for its defining equation. Nontransverse events do not satisfy these conclusions automatically.

For a fixed evaluation input x, the same decomposition reads

$$\partial_{y_i} f_{\theta_\epsilon(t_\epsilon; y)}(x)|_{y_0} = L v_{x,\epsilon} q_i + K_{x,\epsilon,i}.$$

Thus leading label-attribution magnitudes rank examples by $|(XX^\top y_0)_i|$ at every evaluation input with nonzero limiting prediction velocity, provided the compared scores have a strict gap. A finite-scale sufficient inequality is $L|v_{x,\epsilon}|(|q_i| - |q_j|) > |K_{x,\epsilon,i}| + |K_{x,\epsilon,j}|$. At an input with zero limiting velocity, the content can determine the ranking. This is label influence along a training trajectory, not a statement about deleting training examples or converged-estimator influence.

F Derivative-preserving trajectory forecasts

The exact identity (30) separates a computable phase shift from the change in the own-clock output map. The rate forecast and its derivative-preserving version are

$$C_\epsilon^{\mathrm{rate}}(u) = W_\epsilon(y_0, \tau + d_L(u)), \qquad C_\epsilon^{\mathrm{rate}+}(u) = C_\epsilon^{\mathrm{rate}}(u) + K_\epsilon u. \tag{43}$$

They require the nominal trajectory at the shifted time. The second forecast has exact value and derivative at zero. On a compact label-phase domain containing the shifted phase, let $M_y = \sup \|D_y W_\epsilon\|$, $M_{yy} = \sup \|D_y^2 W_\epsilon\|$, and $M_{y\tau} = \sup \|\partial_\tau D_y W_\epsilon\|$; these are bounded uniformly in L. Direct integration in labels gives

$$\|F_\epsilon(t_\epsilon; y_0 + u) - C_\epsilon^{\mathrm{rate}}(u)\| \le M_y \|u\|, \tag{44}$$

$$\|F_\epsilon(t_\epsilon; y_0 + u) - C_\epsilon^{\mathrm{rate}+}(u)\| \le \tfrac{1}{2} M_{yy} \|u\|^2 + M_{y\tau} |d_L(u)| \|u\|. \tag{45}$$

There is no additive finite-initialization error in these inequalities. At $u = v/L$ for bounded v, they are $O(1/L)$ while the true change can be order one. The constants involve nearby-label trajectories, so these inequalities are not an accuracy certificate obtainable from one nominal trajectory alone.

A forecast that does not insert the analytic rate uses $\beta_\epsilon = A_\epsilon^\top v_\epsilon / \|v_\epsilon\|^2$ and $B_\epsilon = A_\epsilon - v_\epsilon \beta_\epsilon^\top$. Define

$$C_\epsilon^{\mathrm{proj}}(u) = F_\epsilon(t_\epsilon + \beta_\epsilon^\top u; y_0) + B_\epsilon u. \tag{46}$$

If the nominal-time argument is admissible, this also has derivative A_ϵ at zero. At fixed emergence phase, $\beta_\epsilon / L \to q$ and $B_\epsilon = O(1)$, so it converges to the same critical random-phase profile for bounded rescaled perturbations. Under gradient flow the training-output envelope extends this conclusion to Gaussian critical noise, by first restricting Z to a compact set and then controlling the tails. If a shifted time is negative, clipping it to zero does not alter the limit: for any fixed rescaled perturbation the shifted time is eventually positive, while the envelope controls the excluded Gaussian tail.

These conclusions are local in phase. When velocity collapses, β_ϵ can be large or undefined even though A_ϵ remains bounded. Small orthogonal-energy fraction alone is not a sufficient forecast certificate: it does not bound content parallel to velocity, extrapolation distance, or trajectory curvature. The rate forecast and the projected forecast therefore need not have the same finite-scale behavior on a plateau.

For GD, the theoretical shift replaces λ by r_h and is measured in steps. Its derivative uses the real-phase derivative in (35). A numerical interpolation convention must use matching derivatives in both the projection and interpolant to retain exact first-order consistency. A finite step increment and a real-phase derivative are different objects. The numerical audit in the experiments tests the implemented convention rather than assuming these objects coincide.

G AN EXPLICIT SCALAR INSTANCE

For $p = ab$ trained on $(p - y)^2/2$ from $(a, b) = (\sqrt{\delta}, 0)$, the invariant $a^2 - b^2 = \delta$ gives

$$\dot{p} = (y - p)\sqrt{\delta^2 + 4p^2}, \qquad |p(t; y)| \le |y|.$$

Let $p(t_\delta; 1) = 1/2$ and $L_\delta = \log(4/\delta)$. Differentiating the hitting-time integral yields

$$J_\delta := \partial_y p(t_\delta; 1) = \frac{\sqrt{1 + \delta^2}}{2} \int_0^{1/2} \frac{du}{(1 - u)^2 \sqrt{\delta^2 + 4u^2}} = \frac{L_\delta + 1}{4} + o(1).$$

For the last equality, the singular part is $\frac{1}{2} \operatorname{arsinh}(1/\delta)$; the remainder converges to $\int_0^{1/2} [(1-u)^{-2} - 1]/(2u)\, du = (\log 2 + 1)/2$.

A direct solution gives the finite-scale profile. Set $d = \delta/2$, $k = (y^2 + d^2)^{1/2}$, $\rho = k + y$, $\zeta = k - y$, and $C = \rho(\rho - d)/(d(\rho + d))$. With

$$z(t) = \frac{\rho - C\zeta e^{-2kt}}{d(1 + Ce^{-2kt})}, \qquad p(t; y) = \frac{d}{2}(z - z^{-1}),$$

the Riccati equation $\dot{z} = 2yz - d(z^2 - 1)$ and $z(0) = 1$ verify the formula. Uniformly near $y = 1$ and for $t = O(L_\delta)$, $p(t; y) = y\ell(2yt - L_\delta - \log y) + O(\delta + \delta^2 L_\delta)$, where $\ell(a) = (1 + e^{-a})^{-1}$. The same estimate holds after a rescaled label derivative. Hence $2t_\delta = L_\delta + o(1)$ and

$$p(t_\delta + u; 1 + v/L_\delta) = (1 + v/L_\delta)\ell(v + 2u + 2uv/L_\delta - \log(1 + v/L_\delta)) + O(\delta + \delta^2 L_\delta).$$

At $u = 0$ this converges in compact C^1 to $\ell(v)$, so Appendix B gives the logistic-normal law and the sharp scale $\sigma L_\delta \to 0$. The finite-scale expression explains why the limiting logistic curve should not be substituted uncritically at modest L_δ.

H SEQUENTIAL TRANSITIONS WITH A SIGNED DESIGN GRAM MATRIX

The nominal saddle-to-saddle limit is studied by Pesme & Flammarion (2023). We restrict to a positive, strictly ordered active-set chain and establish the local label dependence needed for its response law; this does not cover the deactivations allowed in that broader nominal analysis.

This section concerns a distinct solvable model, not arbitrary later transitions of an MLP. Let $X \in \mathbb{R}^{n\times m}$ have full column rank, $G = X^\top X/n \succ 0$, $r = X^\top y/n$, and consider

$$\dot{u} = 2u \odot (r - Gu), \qquad u(0) = \epsilon\mathbf{1}, \qquad F = Xu. \tag{47}$$

Entries of G may have either sign. Coordinates remain positive. If $S = \sum_j u_j$, then $\dot{S} \le 2\|r\|_\infty S - 2\lambda_{\min}(G)S^2/m$, which gives a global bound and rules out finite-time blowup.

For an active set S, use a subscript to denote restrictions and define

$$\begin{aligned} c_j^S &= G_{SS}^{-1}G_{Sj}, \qquad \rho_j^S = 2(r_j - (c_j^S)^\top r_S), \\ \Gamma^S &= G_{S^cS^c} - G_{S^cS}G_{SS}^{-1}G_{SS^c}, \\ u_S^S &= G_{SS}^{-1}r_S, \qquad u_{S^c}^S = 0. \end{aligned}$$

Empty-set quantities have their evident zero-dimensional meanings. After relabeling coordinates, assume the chain $S_k = \{1,\dots,k\}$ has two strict properties at r_0: each plateau u^{S_k} is positive on its active coordinates, and coordinate k uniquely minimizes

$$\tau_j^{S_{k-1}} = \frac{1 - \mathbf{1}^\top c_j^{S_{k-1}}}{\rho_j^{S_{k-1}}} \tag{48}$$

among inactive coordinates with positive $\rho_j^{S_{k-1}}$, with $\rho_k^{S_{k-1}} > 0$. These conditions exclude ties and deactivations. All ensuing estimates are local in labels within this fixed chain; their constants deteriorate near its boundary.

H.1 AN IDENTITY THAT RETAINS THE HISTORY OF EARLIER TRANSITIONS

Since $(\log u)' = 2(r - Gu)$, exact elimination of active columns gives

$$\log u_j(t) - (c_j^S)^\top \log u_S(t) = -(1 - \mathbf{1}^\top c_j^S)L + \rho_j^S t - 2\sum_{i\notin S}\Gamma_{ji}^S \int_0^t u_i(a)\,da. \tag{49}$$

No sign assumption on off-diagonal entries was used. The leading clocks in (48) incorporate earlier plateaus through this identity. To check this explicitly, put $z_j^S(a) = -(1 - \mathbf{1}^\top c_j^S) + \rho_j^S a$. When i joins S, partial regression gives $z_j^{S\cup\{i\}} = z_j^S - (\Gamma_{ij}^S/\Gamma_{ii}^S)z_i^S$. Thus the old and new inactive-coordinate lines agree at the zero of z_i^S. Beginning at $z_j(0) = -1$, induction shows that z_j is continuous with slope $\rho_j^{S_k}$ on plateau k, and its next zero is precisely (48). The strict selection condition makes successive times increase and keeps all other inactive coordinates below zero.

For transition k, let $S = S_{k-1}$, $\rho = \rho_k^S$, $c = c_k^S$, and set

$$\tau_k = \frac{1 - \mathbf{1}^\top c}{\rho}, \quad \chi_k = -\frac{c^\top \log u_S^S}{\rho}, \quad T_k(r) = L\tau_k(r) + \chi_k(r). \tag{50}$$

The offset specifies the incoming amplitude of the transition, rather than a particular arbitrary loss threshold.

Proposition 5 (Sequential entrance and response)**.** *Under the strict chain conditions, there is a bounded incoming profile Ψ_k^r on the face S_k, with $\Psi_{k,S}^r(s) \to u_S^S$ and $e^{-\rho s}\Psi_{k,k}^r(s) \to 1$ as $s \to -\infty$, such that*

$$u_\epsilon(T_k(r) + s; r) \to \Psi_k^r(s)$$

locally in C^1 in (r,s) at rate $O(\epsilon^\beta)$ for some $\beta > 0$. On compact subintervals between successive rescaled transition times, $u_\epsilon(La; r) \to u^{S_k}(r)$ and its first label derivative converges to that of u^{S_k}. At transition k,

$$D_yF = LX(\Psi_k^{r_0})'(s)(q_k^y)^\top + O(1), \qquad q_k^y = \frac{\tau_k(I - P_S)x_k}{x_k^\top(I - P_S)y_0}, \quad P_S = X_S(X_S^\top X_S)^{-1}X_S^\top. \tag{51}$$

In particular the response is asymptotically rank one. On the following plateau it instead converges to P_{S_k}.

Proof. We supply the transition construction and the uniform derivative argument. They are needed because an undifferentiated saddle-chain limit does not suffice. On the face $S\cup\{k\}$ put $b = \log \Psi_S - \log u_S^S$ and $I(s) = \int_{-\infty}^{s} \Psi_k(a)da$. Equation (49) becomes

$$\Psi_k(s) = \exp(c^\top b(s) + \rho s - 2\Gamma_{kk}^S I(s)).$$

The active log equation is $b' = -Mb + N(b) - 2G_{Sk}\Psi_k$, where $M = 2G_{SS}\operatorname{diag}(u_S^S)$ is positive stable and $N(b) = O(\|b\|^2)$. Write its variation-of-constants formula from $-\infty$, together with the integral equation for I. On $s \le s_*$ and in norm $\sup e^{-\omega s}(\|b\| + |I|)$ with $0 < \omega < \min\{\rho, \min \operatorname{Re}\operatorname{spec} M\}$, both equations form a contraction for negative enough s_*. For the first transition S is empty, so the stable equation and its spectral restriction are absent and it suffices to take $0 < \omega < \rho$. Differentiating parameters introduces only factors $1 + |s|$ in $e^{\rho s}$ and polynomial factors in the exponentially decaying stable kernel. Shrinking ω makes their weighted integrals finite. This proves uniqueness and C^1 parameter dependence of the profile.

The profile is global and tends exponentially to u^{S_k}. Indeed, on this face the function $V(u) = \sum_{i\in S_k}(u_i - u_i^{S_k}\log u_i)$ is proper in the positive orthant and satisfies $\dot V = -2(u-u^{S_k})^\top G_{S_kS_k}(u-u^{S_k})$. The only invariant zero-derivative set is the equilibrium. Its linearization is similar to a negative-definite matrix, so convergence is exponential once the profile enters a fixed neighborhood. Compactness of the label neighborhood makes this uniform; differentiating the equation shows the derivative error is bounded by $C(1+s)e^{-cs}$, which is also exponentially small after decreasing c.

The stability estimate used to compare a finite path with this profile does not require nonnegative couplings. If ζ and $\bar\zeta$ are active log coordinates, their difference obeys $d' = -2G_{SS}(e^{\zeta} - e^{\bar\zeta}) + f$. In norm $\|d\|_{G^{-1}}^2 = d^\top G_{SS}^{-1} d$,

$$\tfrac{1}{2}(\|d\|_{G^{-1}}^2)' = -2\sum_i d_i(e^{\zeta_i} - e^{\bar\zeta_i}) + d^\top G_{SS}^{-1} f \le \|d\|_{G^{-1}}\|f\|_{G^{-1}}.$$

The same estimate holds for the linearized equation $a' = -2G_{SS}\operatorname{diag}(w)a + f$ with $w \ge 0$. In particular, the global rate derivative $D_r \log u$ has norm at most Ct, since it solves $(D_r \log u)' = 2I - 2G\operatorname{diag}(u)D_r\log u$.

Choose a small fraction $\nu > 0$ of the minimum clock gap. On a transition window $[T_k - \nu L, T_k + s_1]$, the candidate piecewise-linear logarithmic profiles constructed above are at most $-4b < 0$ for all nonactive coordinates other than k, uniformly in labels. Bootstrap the bound $u_j \le \epsilon^b$ for these coordinates. Formula (49) then improves it to $u_j \le C\epsilon^{4b}$: their own diagonal integral has a nonpositive sign, other still-small coordinates contribute at most $CL\epsilon^b$, and the emerging coordinate contributes a bounded integral on this window. The last bound is bootstrapped simultaneously with closeness of that coordinate to its profile below. Negative Schur off-diagonal entries are allowed because this bounded contribution need not have either sign. Their own-clock derivatives are at most CLu_j by the preceding global derivative bound.

At the start of the window, the previous plateau has approached u^S exponentially over a time proportional to L, and all future-coordinate integrals are exponentially small. The first transition has the same properties with S empty. Compare the active log variables and the integral I with the profile. Their equations are identical except for exponentially small inactive forcing. If $D(s)$ is the sum of their differences, nonexpansion and the explicit factor $e^{\rho s}$ give

$$D(s) \le D(-\nu L) + C\epsilon^b L + C\int_{-\nu L}^{s} e^{\rho a} D(a)da.$$

The integral coefficient is bounded on $(-\infty, s_1]$, so $D = O(\epsilon^{b'})$ for some $b' > 0$. Differentiation at fixed own-clock phase yields the same linear nonexpansive term. Terms involving the already-bounded state error gain only polynomial factors in L, while derivatives of the exponential factor contribute $|s|e^{\rho s}$, still integrable. The derivative difference is therefore $O(\epsilon^{b''})$ after reducing the exponent. On the subsequent plateau, the same nonexpansion estimate prevents amplification over its $O(L)$ length; forcing integrates to a polynomial in L times an exponentially small error. The

profile's exponential convergence then supplies the entrance data for transition $k+1$. Finite induction proves the claimed compact C^1 limits, with one positive exponent after finitely many reductions.

Finally differentiate $u_\epsilon(T_k(r_0)+s;r) = \mathcal{W}_{\epsilon,k}(r, s+T_k(r_0)-T_k(r))$, where $\mathcal{W}_{\epsilon,k}(r,a) = u_\epsilon(T_k(r)+a;r)$. The coefficient of L is $-(\Psi_k)'\nabla_r\tau_k^\top$. Since $-\nabla_r\tau_k = (2\tau_k/\rho)(e_k-\widetilde{c})$, composition with $r = X^\top y/n$ gives (51). The bounded term is $X\{D_r\Psi_k - (\Psi_k)'\nabla_r\chi_k^\top\}X^\top/n$. On a plateau the derivative of $Xu^{S_k}(X^\top y/n)$ is exactly P_{S_k}. □

The profile is nonconstant and X is injective, so $X\Psi_k'(s) \neq 0$ at every finite transition phase. The logistic bound on $\sum u_j$ gives a label-polynomial envelope. Appendix B consequently applies at every declared transition whenever $(q_k^y)^\top \Sigma q_k^y > 0$: the critical scale is $1/L$ and the limiting law is $X\Psi_k(s + c(q_k^y)^\top Z)$. This extension excludes deactivation, nearly tied clocks with vanishing gaps, and arbitrary later MLP transitions. It nevertheless permits signed Gram matrices and gives an analytic timing direction, the feature residualized against the currently active design.

I Experimental protocol and complete comparisons

I.1 Scope, data, and replication

The controlled experiments study first emergence in small-initialization networks. They are mechanism tests, not benchmarks of predictive accuracy or throughput. At $L = 5, 10, 20$, we use five paired initialization seeds and four configurations: gradient flow without biases at $n = 32$ (B0), GD without biases at $n = 32$ (B1), GD with trained biases at $n = 32$ (B2), and GD with trained biases at $n = 128$ (B3). The prospective test at $L = 40$ uses B1–B3, with 15 nominal trajectories and 1,200 perturbed-label retrainings, plus 60 numerical-audit retrainings. The data realization and initialization directions are shared across scales. Thus $L = 40$ is held out in initialization *scale*, not in dataset or initialization seed. The same label-perturbation bank is used across methods and scales to make comparisons paired.

Let $X_{128} \in \mathbb{R}^{128\times 5}$ have independent standard Gaussian entries, generated with NumPy PCG64 seed 742901. Before normalization, the deterministic regression target is

$$y_i^{\mathrm{raw}} = x_{i1} + \tfrac{1}{2}\sin(x_{i2}) + \tfrac{1}{4}x_{i3}x_{i4}.$$

Subtract the mean of the first 32 targets and divide by their population standard deviation. Use this same transformation for all 128 targets, and take the first 32 rows as the smaller dataset. There is no added observation noise in the nominal labels; the Gaussian label perturbations are the experimental intervention. The evaluation set contains 256 independent standard Gaussian inputs, generated with PCG64 seed 742902. It is fixed across all runs.

The model is

$$f_\theta(x) = \sum_{j=1}^{64} a_j \tanh(w_j^\top x + b_j) + b_0.$$

In B1, $b_j = b_0 = 0$ are not trained, leaving 384 parameters. B2/B3 train all 449 parameters. For each seed in $\{201, 202, 203, 204, 205\}$, draw the 384 non-bias coordinates independently from $N(0, 1/64)$ using PCG64 and multiply them by e^{-L}. Initialize all trained biases at zero. The non-bias draw is paired across branches and scales. At $L = 40$, $e^{-L} \simeq 4.25 \times 10^{-18}$; this intentionally extreme initialization isolates the asymptotic regime. It is not proposed as a practical initialization rule.

I.2 Updates, checkpoints, and perturbations

Training minimizes $\ell(\theta; y) = \|F(\theta) - y\|^2/(2n)$ with full-batch GD and $h = 0.05$. Define $g_n = X_n^\top y_n/n$, $\lambda_n = \|g_n\|$, and $r_n = \log(1 + h\lambda_n)$. For reference, $\lambda_{32} = 0.9153977330$ and $\lambda_{128} = 1.2239763954$. The extension fixes requested phases $u \in \{-1, 0, 1, 2, 3, 4, 8\}$ and integer checkpoints

$$k(u) = \left\lfloor \frac{L+u}{r_n} + \frac{1}{2} \right\rfloor.$$

The primary checkpoint is $u = 0$. All perturbed trajectories are evaluated at these nominal checkpoints; their horizons are not changed in response to their losses or peaks. Dense nominal geometry is saved at every integer update spanning reference phase $[-2, 8]$. Figure 3 connects checkpoint measurements for readability; the connecting lines are not additional observations.

The 16 perturbation directions in $\mathbb{R}^{32}$ are independent standard Gaussian vectors from PCG64 seed 742904. For the 128-label branch, append 96 Gaussian coordinates from seed 742905, preserving the first 32 coordinates. Set

$$q_{32} = \frac{X_{32} g_{32}}{32\lambda_{32}^3}, \qquad \|q_{32}\| = 0.2122032201, \qquad \sigma_c = \frac{c}{L\lambda_{32}\|q_{32}\|}.$$

All branches use these same absolute noise scales. Critical levels are $c \in \{0.1, 0.3, 1, 3\}$; the subcritical level is $c = 0.3/L$, equal to 0.0075 at $L = 40$. Each cell therefore has $5 \times 16 = 80$ scientific retrainings. The primary win predictions were specified for $c = 0.3, 1$, with other levels retained as additional comparisons. Perturbed labels are $y + \sigma_c z$ and are never restandardized. Each retraining starts from the same parameters as its nominal run.

I.3 Full response matrices and numerical checks

Let $J_k = \partial_\theta F(\theta_k)$ be the training-output Jacobian and $H_k = \nabla_\theta^2 \ell(\theta_k; y)$. The full parameter-by-label response S_k and the propagated GN approximation N_k start at zero and satisfy

$$S_{k+1} = S_k - hH_k S_k + \frac{h}{n} J_k^\top, \tag{52}$$

$$N_{k+1} = N_k - \frac{h}{n} J_k^\top J_k N_k + \frac{h}{n} J_k^\top. \tag{53}$$

Fresh-output responses are $A_k = J_k^{\mathrm{eval}} S_k$ and $A_k^{\mathrm{GN}} = J_k^{\mathrm{eval}} N_k$. The residual coefficients are held fixed when applying the residual-Hessian term. All trained bias derivatives are included. The 16 perturbations are used for nonlinear retraining comparisons, not as substitutes for the full Jacobian.

The initialization tangent uses every trainable parameter. For initialization-kernel eigenvalue κ, its discrete gain is $[1 - (1 - h\kappa)^k]/\kappa$, with zero-eigenvalue limit hk. Positive eigenvalues are retained even at $L = 40$, using stable `log1p`/`expm1` evaluation. Unresolved tangent outputs are assessed using absolute errors; their relative accuracy is not determined.

The $L = 40$ runs and numerical diagnostics use CPU float64. Detailed numerical checks use seed 201 in each branch; the remaining seeds retain finite-value and completion checks. We compare the exact sensitivity against centered label-retraining finite differences, check residual-Hessian decompositions against the full Hessian action, and compare the spectral tangent with its recurrence. Forecast derivatives are checked separately from training derivatives. Resolved matrix comparisons use relative tolerance 10^{-8}; label and forecast finite differences must meet 10^{-3} at two adjacent step sizes. Quantities below $10^{-10}\sqrt{\text{output dimension}}$ are marked unresolved for relative comparisons. All three branches meet these prespecified criteria. The main experiment took 74 seconds on four CPU workers with one BLAS thread each; numerical validation and pilot runs took 59 and 38 seconds, respectively. These timings describe that implementation and machine, not a hardware-independent complexity guarantee.

I.4 Clock definitions and interpolation

Primary geometry uses the per-update output slope

$$w_k = J_k^{\mathrm{eval}}[-h\nabla_\theta \ell(\theta_k; y)].$$

This is an algorithmic slope convention. It is not asserted to equal the derivative of the theorem's exact real-phase GD orbit. The nominal output interpolant H_0 is cubic Hermite interpolation of integer-update outputs, with these same slopes at the knots. For $0 \le t \le 1$,

$$\begin{aligned} H_0(j+t) = {} & (2t^3 - 3t^2 + 1)F_j + (t^3 - 2t^2 + t)w_j \\ & + (-2t^3 + 3t^2)F_{j+1} + (t^3 - t^2)w_{j+1}. \end{aligned}$$

The projected method uses $\beta = A^\top w/\|w\|^2$ and

$$P_{\mathrm{proj}}(\Delta y) = H_0(k + \beta^\top \Delta y) - H_0(k) + (A - w\beta^\top)\Delta y.$$

The rate method uses

$$q_h = \frac{\nabla_y r(y_0)}{r(y_0)^2}, \qquad d_L(\Delta y) = L\left\{\frac{1}{r(y_0)} - \frac{1}{r(y_0 + \Delta y)}\right\},$$
$$P_{\mathrm{rate}}(\Delta y) = H_0(k + d_L(\Delta y)) - H_0(k) + (A - Lwq_h^\top)\Delta y.$$

Both have derivative A at zero because the same knot slope appears in the interpolation and the linear residual. Each output set uses its own A and w. No shift is fitted to perturbed outputs.

For B1, $r(y) = \log[1 + h\|X^\top y/n\|]$. For B2/B3, define $P = I - \mathbf{1}\mathbf{1}^\top/n$ and $\widetilde{X} = [X, \mathbf{1}]$, and use

$$\widetilde{g} = \widetilde{X}^\top P y/n, \quad \widetilde{\lambda} = \|\widetilde{g}\|, \quad r(y) = \log(1 + h\widetilde{\lambda}), \quad \nabla_y \widetilde{\lambda} = \frac{P\widetilde{X}\widetilde{g}}{n\widetilde{\lambda}}.$$

This analytically centers only the proposed clock. It does not modify the training labels or establish a theorem for biased networks. Geometry-transfer plots use each branch's own phase $kr(y_0) - L$; evaluation checkpoints use the reference rate r_n defined above. In B3 these two coordinates differ slightly, so a requested phase is not relabeled as an exact centered phase.

Negative interpolation queries are clamped to initialization and counted. Positive queries are covered by continuing only the nominal outputs, up to $\lceil (L + 50)/r_n \rceil$ updates. Queries beyond this fixed cap are unavailable, not extrapolated or clamped to the endpoint. A comparison requires coverage of all 16 perturbation directions. Each projected variant has 50 unavailable groups out of 1,050 seed–checkpoint–noise–output-set groups: phase 8 at $c = 3$ in all branches, and phase 8 at $c = 1$ in B1/B2. The rate correction and other baselines have complete coverage. Retraining itself completed in all groups.

I.5 Losses, baselines, and uncertainty

For each seed and condition, with actual response $D_b = F_k(y + \sigma z_b) - F_k(y)$ and forecast P_b, report

$$E(P) = \left\{\frac{\sum_{b=1}^{16} \|D_b - P_b\|^2}{\sum_{b=1}^{16} \|D_b\|^2}\right\}^{1/2}.$$

This is relative RMS error, not relative squared error or an average of direction-wise ratios. We retain numerator and denominator sums as well as absolute errors. Zero change has $E = 1$ whenever the actual response is resolved.

The shrinkage baseline predicts $\gamma_{-s} A_s \Delta y$, where

$$\gamma_{-s} = \mathrm{clip}_{[0,1]} \frac{\sum_{s' \neq s, b} \langle D_{s'b}, A_{s'} \Delta y_b \rangle}{\sum_{s' \neq s, b} \|A_{s'} \Delta y_b\|^2}.$$

Fitting uses the other four initialization seeds at the same branch, phase, noise level, and output set. It uses additional retraining information unavailable to the clock methods, but not the target seed's outcomes. It is therefore a strong supervised calibration comparator, rather than a compute-matched method. A tie is declared when error differences are at most $10^{-10} \max(1, E_1, E_2)$; unavailable entries are kept separate.

Lines in the figures are medians over the five paired seeds. Bands span their observed values and are not confidence intervals. Directions, phases, branches with nested data, and paired L values are not independent statistical replications. We do not derive inferential significance from the total number of paired wins.

I.6 Prospective geometry test

The scale-transfer curves were fixed using $L = 20$ geometry before any $L = 40$ result was generated. Positive scalar norms are interpolated log-linearly on branch phase, without extrapolation.

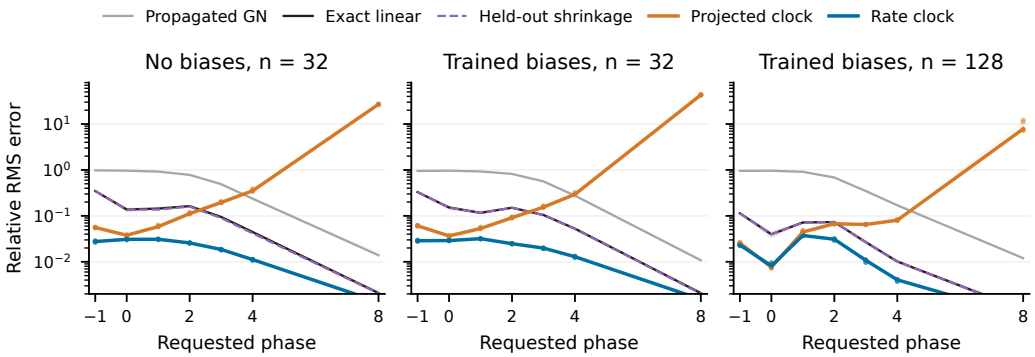


Figure 4: **Forecasts at the other primary noise level, $c = 0.3$.** Conventions match Figure 3. All methods have complete coverage at this noise level. The rate correction retains an advantage across phases, whereas the projected correction becomes inaccurate on the plateau.

Equation (22) is evaluated from the source-run perpendicular norm, parallel norm, velocity norm, and analytic q_h. No measured $L = 40$ quantity enters the forecast.

Transfer is assessed on $[-1, 4]$ in increments of 0.05. Log-scale R^2 is computed across the five paired seed curves as a descriptive goodness-of-prediction measure, not as an independence-based statistic. Median absolute $\log_{10}$ errors are 0.00760, 0.00453, and 0.00991 for B1–B3. The no-change baseline $\eta_{40} = \eta_{20}$ gives $R^2 = 0.820, 0.618, 0.727$; the asymptotic rescaling $\eta_{20}/4$ gives $0.934, 0.877, 0.846$. The source-only forecast gives $0.999806, 0.999697, 0.999758$.

The window boundary is the first post-speed-peak grid point where median η exceeds 0.075. Its numerical agreement tests a fixed descriptive threshold, not a certified boundary of forecast validity. Operator-norm peaks lag speed peaks by median phases $0.0424, 0.0429, 0.0502$, within the prespecified interval $(0, 0.06]$. These peak locations use three-point parabolic interpolation on the fixed grid and are not exact continuous maxima. The predicted 1% bound on all source-to-target perpendicular-norm changes fails slightly in B1/B2; no tested point exceeds 5%. Norm agreement cannot establish that full matrices, their orientations, or their parallel components are identical.

I.7 Complete noise sweep and failures

The full sweep in Table 2 shows why a win statement must specify the noise level. At $c = 0.3$ and 1, rate forecasts beat linear influence, GN, and shrinkage at every tested phase. At $c = 3$, the same ordering holds, although absolute errors are larger. At $c = 0.1$, the rate method loses to linear influence in 25 of 105 comparisons. At the subcritical level it loses in 92 of 105 comparisons. In that regime the exact derivative already has small error; preserving the derivative does not ensure a smaller finite-noise remainder.

For the two primary levels, all 120 comparisons through phase 2 favor the projected method over each of linear influence, GN, and shrinkage. At phases 3 and 4, all 60 comparisons favor linear influence and shrinkage over the projected method. At phase 8, all 20 available comparisons do so, with ten missing. The rate method has 210 wins against each of these three baselines, but its plateau benefit is much smaller than its transition benefit: the median paired rate/linear error ratio is 0.207 at phase 0 and 0.712 at phase 8. The median paired absolute error reduction at phase 8 is 7.93×10^{-4}.

Small η is a local statement about derivative geometry. It neither bounds the curvature of the trajectory over a finite shift nor ensures that a projected shift stays in the well-resolved part of the nominal trajectory. Conversely, the rate forecast need not become inaccurate whenever η rises: its shift is computed from a label-dependent rate rather than by dividing by the instantaneous speed. These distinctions explain why the two implementations separate on the plateau without turning the observed threshold into a universal decision rule.

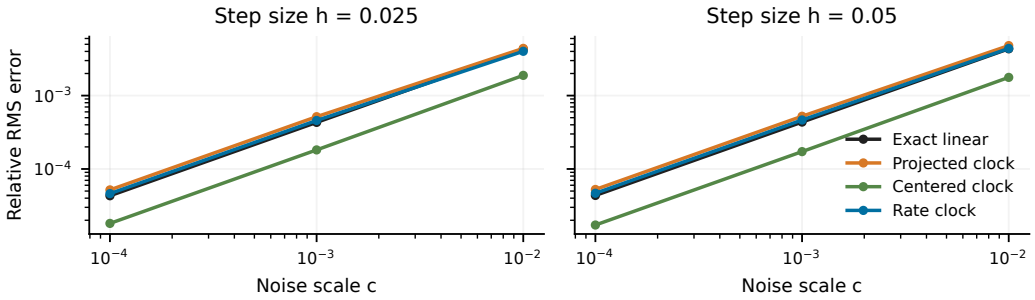


Figure 5: **Small-noise validation of forecast derivatives.** Medians pool the fixed phases, two directions, and two signs within one diagnostic seed; they do not quantify uncertainty across initialization seeds. Error decreases at smaller noise for each slope convention. The centered convention is a distinct interpolation approximation, not the exact real-phase GD derivative.

I.8 The finite-step derivative audit

We check local derivative consistency separately from finite-perturbation accuracy. At an integer checkpoint, the Hermite interpolant's derivative is its specified knot slope. Using the same slope in the projected clock and linear residual gives the chain-rule derivative A. We check this identity numerically using the response matrix, nominal anchor, and interpolation knots for each output set.

The separate small-noise diagnostic uses B1, seed 201, $L = 20$, step sizes 0.05 and 0.025, the first two directions with both signs, and $c = 10^{-4}, 10^{-3}, 10^{-2}$, evaluated at requested phases $-1, 0, 1$. This diagnostic uses 24 of the numerical-audit retrainings, not additional scientific replicates. At $h = 0.05$, the projected-update median errors are 5.26×10^{-5}, 5.23×10^{-4}, and 4.80×10^{-3}, respectively, decreasing with noise over the tested range. A centered-slope Hermite variant uses $(F_{j+1} - F_{j-1})/2$ at interior knots and $F_1 - F_0$ at initialization, with the same slope in the projected clock and residual. It also preserves the derivative A and gives smaller median errors in this diagnostic. Halving the step size does not uniformly improve every convention.

These measurements are consistent with finite-noise and finite-shift interpolation contributions to forecast error. They do not uniquely separate those contributions, prove that the Hermite interpolant equals the GD emergence orbit, or establish the full fixed-step asymptotic forecast theorem for this numerical implementation.

I.9 Consistency checks for figures and tables

We independently recomputed the 9,300 finite forecast errors from saved squared-error sums, retained all 150 unavailable entries, and checked all 4,200 paired-summary rows against seed-level errors. We also checked the source-only geometry forecast against Equation (22). Separate checks using saved response and prediction arrays verified forecast calculations, including held-out shrinkage coefficients. Summary recomputation and saved-array checks validate the analysis without rerunning model training. All unavailable entries and subcritical losses are retained.