KERNEL VELOCITY: SEPARATING FEATURE CHANGE FROM PREDICTIVE BENEFIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature changes during neural-network training can alter predictions without improving them. We study this distinction through the residual action of kernel velocity: the derivative of the neural tangent kernel (NTK) applied to the training residual. For smooth models under squared-loss gradient flow with a constant preconditioner, this vector determines the generic quadratic departure from a matched frozen-kernel predictor. An exactly matched two-layer family holds the initial predictor and kernel fixed while changing the departure scale; symmetry identifies cases where a cubic term replaces it. We connect the local result to discrete updates and to an exact risk identity that separates error alignment from prediction-change magnitude. Independent derivative checks validate the formulas. A discrete-training study finds incremental short-horizon information beyond parameterization controls and compares simpler diagnostics. Across 96 controlled settings, the signed contrast predicts early risk ordering correctly but loses accuracy over longer horizons. Matched real-data comparisons exhibit similar feature-change magnitudes with opposite predictive outcomes, and trajectory measurements explain sign reversals through changes in direction and remaining error. Validation-selected and compute-matched controls delimit the resulting practical claims. The method provides a task-specific way to investigate feature learning and interpret frozen-kernel failure; the experiments do not establish a broadly effective training controller or optimizer speedup.

1 INTRODUCTION

The neural tangent kernel describes how parameter updates move a network's predictions (Jacot et al., 2018; Lee et al., 2019). When the kernel stays close to its initial value, a linear model can approximate training. Other parameterizations permit substantial feature adaptation, including at infinite width (Chizat et al., 2019; Mei et al., 2018; Yang & Hu, 2021). This raises two distinct questions: which feature changes affect the predictor, and which of those effects improve prediction?

Neither parameter movement nor a kernel-change norm answers both questions. Kernel motion can occur in directions that do not act on the current residual. Even when that motion changes predictions, its effect on held-out error depends on direction as well as magnitude. Comparing networks with different initial predictors or kernels introduces a further ambiguity: subsequent performance differences may reflect the starting predictor or optimization geometry rather than feature adaptation.

We propose residual-action analysis for this controlled comparison. Given the same initial predictor and effective kernel, the vector $\dot{K}_0 r_0$ determines the generic second-order departure from frozen-kernel training. The result identifies task-visible tangent-feature evolution without equating it with better generalization. We construct an exactly matched finite-width family to isolate the departure scale, characterize symmetry cancellations, and derive a signed held-out contrast. An exact risk decomposition then explains how an initially favorable change can become unfavorable as the prediction difference and the remaining error evolve.

The contribution is the connection between a local predictor-level certificate, exact initialization controls, and an explicit separation of departure from benefit. Section 2 develops the certificate and its finite-time limits. Section 3 supplies the matched construction and a separate solvable alignment model. Section 4 connects the certificate to computation, discrete updates, and risk. Sections 5

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.

and 6 test those connections and their limits. This framework offers an alternative to interpreting feature-change magnitude alone as evidence of useful learning.

Relation to prior work. Higher-order kernel dynamics already appear in the neural tangent hierarchy, an infinite system obtained by repeatedly differentiating the evolving kernel (Huang & Yau, 2020). Dynamical mean-field theory describes evolving kernels in feature-learning limits (Bordelon & Pehlevan, 2022), and adaptive-kernel and multiscale theories connect learned representations to predictors across scaling regimes (Lauditi et al., 2025; Rubin et al., 2025). Our finite-width construction isolates the residual-conditioned departure while matching the entire initial predictor and effective kernel. We do not claim a new kernel hierarchy or a general solution of feature-learning dynamics.

The feature-speed formula relates representation updates to the backward pass and loss decay (Chizat & Netrapalli, 2024). We instead ask how kernel evolution changes a matched output trajectory and its held-out risk. These quantities answer different questions; our experiments test their practical interpretation rather than assert universal diagnostic superiority. Learned tangent kernels and task alignment also predate this work (Long, 2021; Atanasov et al., 2022). We use kernel refreshing as a control. Statistical separations from kernel methods (Ghorbani et al., 2020) and exact linear-network dynamics (Saxe et al., 2014) address broader or different settings than the local comparison studied here.

2 A LOCAL CERTIFICATE FOR FROZEN-KERNEL FAILURE

For inputs $x_i \in \mathbb{R}^d$, labels $y \in \mathbb{R}^N$, and parameters $\theta \in \mathbb{R}^P$, let $f(\theta) = (f(x_i;\theta))_{i=1}^N$, $r = f - y$, and $L = \|r\|^2/(2N)$. A constant symmetric positive semidefinite preconditioner A includes fixed layerwise learning-rate scales. Write $J = \partial f/\partial\theta \in \mathbb{R}^{N\times P}$ and

$$\dot\theta = -AJ^\top r/N, \qquad K = JAJ^\top, \qquad \dot f = -Kr/N. \tag{1}$$

The matched frozen trajectory satisfies $\dot f_t^{\mathrm{K}} = -K_0(f_t^{\mathrm{K}} - y)/N$ and $f_0^{\mathrm{K}} = f_0$. All vector norms are Euclidean; $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ denote Frobenius and spectral norms. A dot denotes differentiation along gradient flow, including the scale in A. We call $\dot K_0 r_0$ the residual-action vector; the normalized rates below are scalar diagnostics.

Proposition 1 (Label-conditioned velocity)**.** *For twice continuously differentiable predictions, let $g_i = \nabla_\theta f(x_i;\theta)$ and $H_i = \nabla_\theta^2 f(x_i;\theta)$. Then*

$$\dot K_{ij} = -\frac{1}{N}\sum_k r_k\left(g_k^\top A H_i A g_j + g_i^\top A H_j A g_k\right). \tag{2}$$

Thus the velocity depends linearly on the instantaneous residual at fixed parameters.

Theorem 2 (Residual-action certificate)**.** *Assume each prediction is C^3 in a neighborhood of a solution on $[0,\tau]$. Then*

$$e_t := f_t - f_t^{\mathrm{K}} = -\frac{t^2}{2N}\dot K_0 r_0 + R_3(t), \qquad \|R_3(t)\| \le C_\tau t^3. \tag{3}$$

If $\dot K_0 r_0 \neq 0$, some $\tau' > 0$ satisfies $\|e_t\| \ge t^2\left\|\dot K_0 r_0\right\|/(4N)$ for $0 < t \le \tau'$.

Both flows have the same initial value and first derivative. Their second derivatives differ by $-\dot K_0 r_0/N$, which proves the leading term. Appendix A gives the proof and an explicit bound for C_τ in terms of derivatives along the trajectory. Initial velocity alone does not determine these bounds.

Proposition 3 (Finite-time bound and cubic fallback)**.** *Let $U(t,s)$ solve $\partial_t U = -K_t U/N$, $U(s,s) = I$. Then*

$$e_t = -\frac{1}{N}\int_0^t U(t,s)(K_s - K_0) r_s^{\mathrm{K}}\, ds, \tag{4}$$

$$\|e_t\| \le \frac{\|r_0\|}{N}\int_0^t \|K_s - K_0\|_{\mathrm{op}}\, ds. \tag{5}$$

Under Theorem 2's assumptions,

$$e_0^{(3)} = -\frac{1}{N}\ddot{K}_0 r_0 + \frac{1}{N^2}(2\dot{K}_0 K_0 + K_0\dot{K}_0)r_0. \tag{6}$$

If $\dot{K}_0 r_0 = 0$ *and* $e_0^{(3)} \neq 0$*, then* $e_t = t^3 e_0^{(3)}/6 + o(t^3)$.

Positive semidefiniteness makes U contractive even when kernels at different times do not commute. Small operator-norm drift is sufficient for accurate frozen prediction; its converse need not hold. Equation (6) also shows why vanishing first velocity cannot establish global laziness.

Normalization and the clock. The geometric and task-effective rates are $\nu_K = \left\|\dot{K}_0\right\|_F / \|K_0\|_F$ and $\nu_{\text{eff}} = \left\|\dot{K}_0 r_0\right\| / \|K_0 r_0\|$. Both have inverse-time units. The leading nonlinear correction divided by the leading frozen displacement is $t\nu_{\text{eff}}/2$; $2/\nu_{\text{eff}}$ is only a Taylor scale, not a guaranteed interval of accuracy. A dimensionless alternative is

$$\chi = \frac{N\left\|\dot{K}_0 r_0\right\| \|r_0\|}{\|K_0 r_0\|^2}, \qquad s = \frac{t\,\|K_0 r_0\|}{N\,\|r_0\|}. \tag{7}$$

The leading displacement ratio is $s\chi/2$. Under $A \mapsto cA$ with $c > 0$, $K_0 \mapsto cK_0$ and $\dot{K}_0 \mapsto c^2\dot{K}_0$, so χ is invariant. This removes a global clock choice, not task dependence. If $K_0 r_0 = 0$, positive semidefiniteness implies $AJ_0^\top r_0 = 0$; uniqueness makes the autonomous trajectory stationary. Thus the zero denominator identifies a stationary initialization, which we exclude from normalized diagnostics.

3 Exactly matched scaling and symmetry

For $\phi \in C^3$, scalar readouts c_a, and hidden weights $w_a \in \mathbb{R}^d$, define the centered two-layer family

$$f_{m,\alpha}(x;\theta) = m^{-\alpha}\sum_{a=1}^{m}[c_a\phi(w_a^\top x) - c_{a,0}\phi(w_{a,0}^\top x)], \quad A = m^{2\alpha-1}I, \quad \alpha \in [1/2, 1]. \tag{8}$$

The subtracted initialization is constant during differentiation. This is a diagnostic construction that matches initial functions, not a claim about arbitrary production networks.

Theorem 4 (Matched kernel and separated velocity)**.** *Every family member has* $f_0 = 0$*. At any common parameter value,*

$$K_{ij} = \frac{1}{m}\sum_a [\phi_{ai}\phi_{aj} + c_a^2\phi'_{ai}\phi'_{aj}x_i^\top x_j], \tag{9}$$

independently of α*, where* $\phi_{ai} = \phi(w_a^\top x_i)$*. At initialization,*

$$\dot{K}_0^{(\alpha)} = m^{\alpha-1}V_m \tag{10}$$

for a matrix V_m *independent of* α*. If nested i.i.d. particles have integrable kernel and velocity summands, with nonzero limits* $K_{m,0} \to K_\infty$ *and* $V_m \to V_\infty$*, then* $\nu_K = \Theta(m^{\alpha-1})$ *almost surely. If additionally* $V_\infty r_0 \neq 0$ *and* $K_\infty r_0 \neq 0$*, the same conclusion holds for* ν_{eff}.

Thus the entire frozen trajectory agrees across α at a given width, while the nonlinear acceleration differs exactly by $m^{\alpha-1}$. The population nondegeneracy assumption is essential. Without centering, the kernel still matches across exponents, but the initial predictor and residual generally change. The velocity law then contains an exponent-dependent residual; Appendix B gives the distinction explicitly.

Proposition 5 (Cancellation)**.** *Suppose the velocity summands have finite second moments and* $K_{m,0} \to K_\infty \neq 0$*. If the i.i.d. particle law is invariant under* $c \mapsto -c$*, or if* ϕ *is odd and the law is invariant under* $w \mapsto -w$*, then* $\nu_K = O_p(m^{\alpha-3/2})$*. The same rate holds for* ν_{eff} *if* $K_\infty r_0 \neq 0$*. Exact pairs* $(c, w), (-c, w)$ *instead give* $\dot{K}_0 = 0$ *and*

$$\ddot{K}_0^{(\alpha)} = m^{2\alpha-2}W_m, \qquad e_0^{(3)} = -m^{2\alpha-2}W_m r_0/N, \tag{11}$$

where W_m *is independent of* α.

Appendix B gives the explicit summands and pairwise argument. The tanh prediction experiments below have population first-velocity cancellation. Their later gains therefore do not establish that the initial second-order norm predicts useful learning.

An exact task-alignment mechanism. As a separate solvable example, consider the population linear teacher $y = u^\top x$, $u \neq 0$, $\mathbb{E}[xx^\top] = I$, and rank-one network $f(x) = aw^\top x$. Euclidean gradient flow conserves the imbalance $\delta = a^2 - \|w\|^2$. Writing $q = aw$ and $\gamma = (\delta + \sqrt{\delta^2 + 4\|q\|^2})/2 = a^2$, the predictor closes as

$$\dot{q} = -\left(\gamma I + \frac{qq^\top}{\gamma}\right)(q - u), \qquad \dot{\rho} = \frac{\gamma\|u\|}{\|q\|}(1 - \rho^2), \quad \rho = \frac{q^\top u}{\|q\|\,\|u\|}. \tag{12}$$

The input-space matrix $M_t = \gamma I + qq^\top/\gamma$ induces the training-set NTK $K_t = XM_tX^\top$, where row i of X is $x_i^\top$. This connects the alignment model to the earlier kernel object. The invariant δ follows from $d(a^2 - \|w\|^2)/dt = 0$; the expression for γ eliminates the factor parameters and closes the dynamics in q. The closure requires $\gamma > 0$, and the angular law requires $q \neq 0$. Signed alignment increases; the unoriented kernel axis $|\rho|$ can initially decrease when $\rho < 0$. For the population loss $L = \|q - u\|^2/2$, positive imbalance gives $L(t) \leq L(0)e^{-2\delta t}$. Appendix C proves this mechanism and states its domain conditions.

4 Computable action and signed held-out risk

Matrix-free evaluation. Let $v_0 = -AJ_0^\top r_0/N$ and hold r_0 fixed when differentiating the action. Then

$$\dot{K}_0 r_0 = D_\theta[J(\theta)AJ(\theta)^\top r_0]_{\theta_0}[v_0] = \dot{J}_0 AJ_0^\top r_0 + J_0 A\dot{J}_0^\top r_0. \tag{13}$$

A vector-Jacobian product (VJP) followed by a Jacobian-vector product (JVP) evaluates the inner kernel action; one directional derivative evaluates its velocity. The routine avoids both an $N \times P$ Jacobian and an $N \times N$ kernel. Our implementation supports constant diagonal A, including the matched family. Repeated Lie derivatives implement the cubic fallback, differentiating the parameter vector field as well as the output. Holding the vector field fixed at every derivative would be incorrect.

A local risk contrast. Norms certify departure without its predictive sign. Let V contain N_V held-out labeled examples, $B_t = J_V(t)AJ_X(t)^\top$, and let frozen held-out predictions evolve with B_0 and the frozen training residual. Neither held-out inputs nor labels enter the training vector field.

Theorem 6 (Signed held-out contrast)**.** *Under local C^3 smoothness on training and held-out examples and matched initial predictions, write $R_V = \|r_V\|^2/N_V$ for full mean squared error (MSE). Then*

$$f_V(t) - f_V^{\mathrm{K}}(t) = -\frac{t^2}{2N}\dot{B}_0 r_{X,0} + O(t^3), \tag{14}$$

$$R_V(t) - R_V^{\mathrm{K}}(t) = -\frac{t^2}{NN_V} r_{V,0}^\top \dot{B}_0 r_{X,0} + O(t^3). \tag{15}$$

A nonzero signed coefficient determines the sufficiently early MSE ordering relative to the frozen model.

This uses held-out labels and compares two predictors; it is not a label-free population guarantee, an improvement over initialization, or a long-time stopping rule. Appendix D proves the statement for general nonzero initialization.

Discrete updates. The flow coefficient does not give the complete finite-step correction. Let k be a fixed number of gradient-descent (GD) updates of size η, let $u_0 = AJ_0^\top r_{X,0}/N$, and define $h_{V,0} = D_\theta^2 f_V(\theta_0)[u_0, u_0]$. Against the matched linearized GD predictor,

$$e_{V,k} = -\frac{\eta^2 k(k-1)}{2N}\dot{B}_0 r_{X,0} + \frac{k\eta^2}{2}h_{V,0} + O(\eta^3). \tag{16}$$

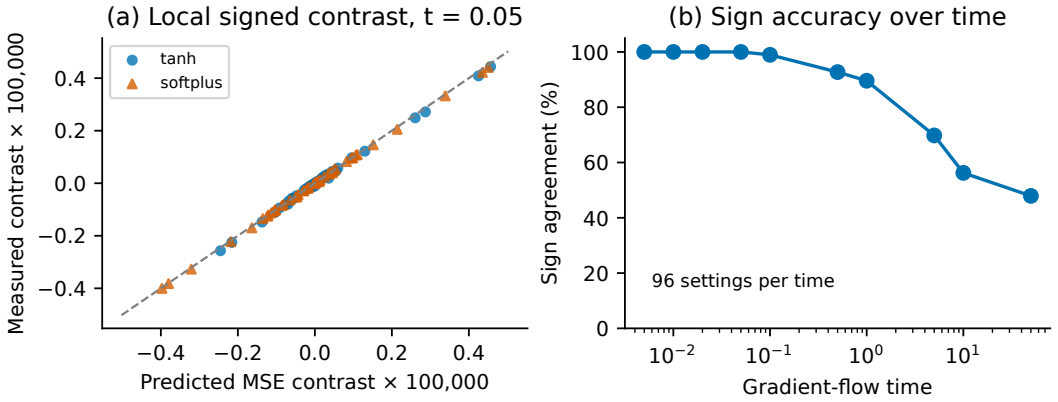


Figure 1: Signed risk is local. Across 96 settings (16 seeds, two activations, three training-label families), the formula predicts both improvement and deterioration at $t = 0.05$. Sign agreement is 100% at all tested times through 0.05, but falls to 47.92% by $t = 50$. These are repeated observations of the same settings, not independent samples at each time.

Here $\eta \to 0$ at fixed k, and $\dot{B}_0$ denotes the gradient-flow derivative. Output curvature supplies the first-step gap; the accumulated kernel-velocity term enters from the second step. Appendix E proves the expansion and reports independent step-halving checks. It does not extend the remainder uniformly to convergence.

From departure to predictive benefit. For any pair of nonlinear and frozen predictions, at any time, let $e_V = f_V - f_V^{\mathrm{K}}$ and $r_V^{\mathrm{K}} = f_V^{\mathrm{K}} - y_V$. The exact identity

$$R_V - R_V^{\mathrm{K}} = \frac{2(r_V^{\mathrm{K}})^\top e_V + \|e_V\|^2}{N_V} \tag{17}$$

separates alignment with the remaining error from the size of the change. When both norms are nonzero, improvement requires $\cos(r_V^{\mathrm{K}}, e_V) < -\|e_V\| / (2\left\|r_V^{\mathrm{K}}\right\|)$. A change may point toward lower error yet overshoot. The local coefficient predicts the initial direction; later rotation of e_V or changes in r_V^{K} can reverse the risk ordering. This elementary identity provides the interpretation of the certificate, rather than an additional generalization theorem.

5 Experiments: Mechanisms, Controls, and Limits

Design and separation of evidence. We separate synthetic mechanism tests, independent derivative checks, validation-selected prediction comparisons, and tests of training decisions. The reference experiments, fixed 64-seed primary comparison, independent precision expansion, and exploratory audits remain separate in all summaries. We select methods using validation error and reserve test error for evaluation. Appendices G–H give the protocols and extended results.

Controlled mechanisms and stopping-time effects. The width-law slopes are $-0.5107, -0.2607, -0.0107$ for $\alpha = 0.5, 0.75, 1$. At width 2048 and common time 50, adaptive test MSE is 0.160437 versus frozen 0.293926, with teacher feature-spike alignment 0.9312. However, the gain-2 frozen test oracle reaches 0.155643. The common-time comparison therefore combines adaptation with acceleration. At gains 3–5 the adaptive network beats that diagnostic oracle by $0.00849, 0.01546, 0.01934$. Test oracles are descriptive controls only. Label rotations and permutations reproduce the task-directed covariance change, without proving that the local norm predicts its usefulness.

The reference width driver obtains different-α velocities by the exact scaling identity. Its shifted fitted slopes are consequently not independent tests of that identity. We address this with 1,536 direct derivative evaluations across 16 seeds, eight widths, three exponents, two activations, and standard/antithetic initializations. Matrix-free and explicit-kernel actions agree to maximum absolute norm discrepancy 9.18×10^{-16}. Fourteen analytic and implementation tests pass. In a separate

Gain	Frozen	Ridge	Adaptive	Refreshed	Paired improvement (95% CI)
2	0.1522	0.1533	0.1124	0.1082	0.0042 [0.0037, 0.0047]
3	0.2340	0.2349	0.1853	0.1794	0.0059 [0.0053, 0.0065]
4	0.2881	0.2890	0.2383	0.2323	0.0060 [0.0053, 0.0068]
5	0.3251	0.3261	0.2766	0.2707	0.0060 [0.0051, 0.0069]

Table 1: Primary study: mean test MSE over 64 paired seeds per gain. Every method selects its checkpoint or penalty using validation MSE. Improvement is adaptive minus refreshed MSE; intervals are paired-seed percentile bootstrap intervals (2,000 resamples).

cubic check, the mean measured-to-predicted norm ratio at $t = 0.01$ is 0.99793, with minimum directional cosine 0.99999984.

Local prediction and its breakdown. The signed-risk study uses $N = 32$, $N_V = 128$, $d = 12$, width 128, tanh/shifted softplus, and teacher, permuted, or norm-matched random training labels. Across 96 settings, all signs are correct at each of $t = 0.005, 0.01, 0.02, 0.05$; there are 52 improvements and 44 deteriorations. All predicted magnitudes exceed the fixed 10^{-11} scoring threshold. Extending the same trajectories gives agreement $98.96\%, 92.71\%, 89.58\%, 69.79\%, 56.25\%, 47.92\%$ at $t = 0.1, 0.5, 1, 5, 10, 50$, respectively (Figure 1). The positive local result and the failure of extrapolation are equally relevant.

Validation-selected prediction study. Inputs follow $x \sim \mathcal{N}(0, I_d/d)$ and labels $y = \tanh(gu^\top x)$ with $\|u\| = \sqrt{d}$. The primary study uses $N = 128$, $d = 50$, width 2048, 256 validation and 2048 test examples, tanh, $c_{a,0} = 1$, and Gaussian hidden weights. Independent random streams generate teacher, training, validation, test, noise, and initialization; seeds are paired across methods and gains $g \in \{2, 3, 4, 5\}$.

We compare four validation-selected methods. *Adaptive* selects a nonlinear checkpoint in $\{5, 10, 25, 50, 100, 200\}$. *Frozen* selects time on an exact initial-kernel flow grid up to 10^6. *Ridge* selects an initial-kernel regularization penalty. *Refreshed* chooses a refresh time in $\{5, 10, 25, 50\}$ and exact frozen continuation from the current nonzero predictor using the current train and cross kernels. Continuation uses the current training residual, not a refit from zero. Its larger selection grid and exact continuation mean this is not a compute-matched algorithm comparison.

Table 1 shows refreshed-over-adaptive improvements of 0.00418–0.00599. Paired one-sided signed-rank tests remain significant after Holm correction across the four gains (adjusted $p \le 7.49 \times 10^{-12}$). The independent expansion completes 555 balanced seeds per gain and gives improvements 0.004118 $[0.003990, 0.004253]$, 0.005594 $[0.005366, 0.005839]$, 0.005890 $[0.005624, 0.006153]$, and 0.005760 $[0.005474, 0.006058]$. The stopping rule is the preset wall-clock budget, not a significance threshold. We retain two additional completed gain cases from an incomplete seed block in the raw data and exclude them from balanced summaries.

Longer training and negative controls. Figure 2 summarizes the horizon dependence. Extending nonlinear training to time 1,000 on 16 paired seeds per gain reduces the gain-2 improvement to 0.000606; many runs still select the boundary. A further four-seed audit to 10,000 gives 0.000661 with interval $[-0.000265, 0.001926]$. Other gains retain positive point estimates, but four seeds cannot establish a general converged-network advantage. The refresh candidates remain the early checkpoints; the audit changes the competing nonlinear horizon.

Eight secondary conditions vary sample count, width, activation, training-label noise, and antithetic initialization, with 16 paired seeds at gains 2 and 4. All refreshed-over-adaptive point differences are positive at horizon 200. Yet in the softplus condition, the initial frozen kernel beats refreshed by 0.00206 and 0.00308, respectively. Noise experiments corrupt training labels while evaluating clean targets, so they measure clean-function recovery. These exploratory comparisons do not establish uniform superiority across tasks or activations.

CIFAR scope and computation. The CIFAR study includes 15 local-certificate and 15 performance task-seed cases. With task-level clustering over only five fixed binary tasks, the validation-

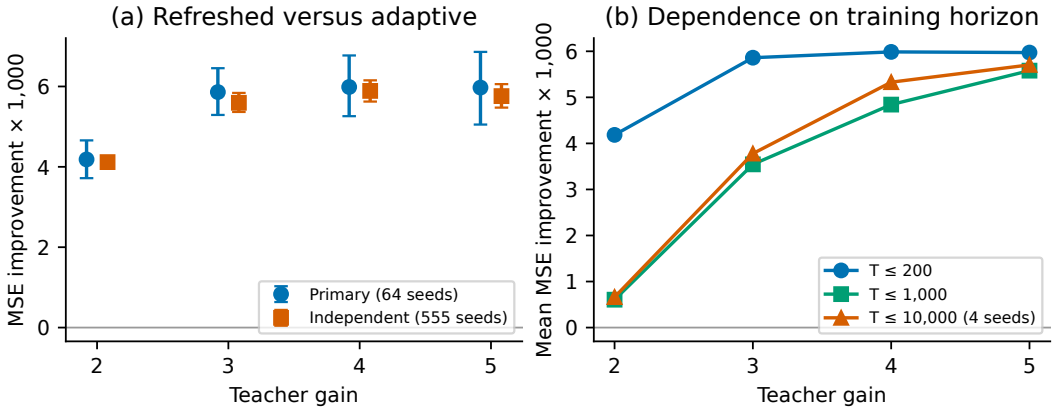


Figure 2: Validation-selected refreshed-kernel benefit and its limits. Left: fixed primary and independent replication estimates with paired 95% bootstrap intervals. Right: longer nonlinear training reduces the advantage, especially at gain 2. Horizon studies reuse primary seeds (16 at 1,000; four at 10,000) and are exploratory; connecting their means does not remove their differing sample sizes.

selected adaptive MSE advantage is 0.003744 with bootstrap interval $[-0.003550, 0.012107]$. The same-time accuracy gain is 2.17 percentage points, but stopping controls do not establish a reliable advantage. That experiment compares the initial kernel only; Section 6 adds refreshed-kernel comparisons on CIFAR.

On a four-thread CPU benchmark with width 256, the matrix-free action takes 4.61 ms at $N = 256$, versus 172.43 ms for explicit autodiff-Jacobian differentiation (about $37\times$; Appendix F). This is an implementation-specific certificate benchmark, not a speedup over an analytic kernel or of the full prediction method. Halving the primary integration step on four paired seeds changes selected adaptive and refreshed test MSE by less than 10^{-9}.

6 Transfer and Training Decisions

A separate test of practical usefulness. We fixed diagnostic rules on calibration tasks before evaluating unseen configurations. The transfer study contains 480 paired benchmarks across Fashion-MNIST, CIFAR-10, wine, diabetes, and two synthetic teacher families, plus 720 timed policy runs on the four real-data families. Small multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) use full-batch or minibatch training. At a common 25-update probe, we measure parameter movement, hidden-representation Gram change, kernel drift, residual-action velocity χ, and the signed validation contrast. Training and validation MSE provide context. Each benchmark compares nonlinear training with initial and refreshed tangent models under the same update stream and selects checkpoints by validation MSE. Here, benefit means refreshed-kernel test MSE minus nonlinear test MSE: it is the negative of the risk gap in Equation (17). Appendix I gives the split, calibration, budgets, and numerical repair.

Similar change, different outcomes. Within dataset, architecture, and batch regime, we pair runs whose change magnitudes differ by at most 10%. Opposite benefit signs occur in 186 of 666 kernel-drift pairs (27.9%) and 254 of 761 representation-change pairs (33.4%). Excluding pairs with either absolute benefit at most 0.001 leaves rates of 26.3% and 30.5%. These overlapping pairs provide descriptive counterexamples to reading predictive benefit from change magnitude alone. They do not establish causal mediation or imply that magnitude carries no statistical information.

Limited transfer beyond the local regime. The raw signed contrast predicts 278 of 480 benefit signs correctly (57.9%; Table 2). Accuracy falls from 60.9% in the primary cohort to 54.5% in the fresh reserve. Its Spearman correlation with benefit magnitude is only 0.086. On the 244 cases with absolute benefit above 0.01, accuracy is 55.3%, versus 67.6% for the context rule. Representation change reaches AUC 0.601, so the evidence does not support dismissing magnitude diagnostics

Diagnostic rule	Sign accuracy (%)	AUC
MSE context only	50.2	0.477
Parameter movement	50.2	0.315
Representation change	50.2	0.601
Kernel drift	49.8	0.465
Residual-action velocity	50.2	0.433
Learned signed contrast	49.8	0.381
Combined diagnostics	48.1	0.474
Raw signed contrast	57.9	0.589

Table 2: Transfer to 480 unseen benchmark cases. Accuracy predicts the sign of nonlinear benefit over the refreshed tangent model; area under the receiver operating characteristic curve (AUC) ranks that binary outcome. Learned rules include training and validation MSE context. The raw signed rule uses the contrast directly without calibration.

altogether. These results show limited long-horizon association, not reliable task transfer. The flow theorem does not cover minibatch updates; at each stochastic probe we evaluate derivatives of the full-batch gradient field.

Incremental information at short horizons. A separate completed study evaluates 55,200 configurations in 345 seed blocks, with centered and ordinary two-layer networks, fixed metrics, and three discrete step scales. Rules fitted on three development tasks predict log prediction-gap RMS on abalone and wine quality after controlling for width, scaling, step size, and initialization. At eight updates, residual action reduces squared forecast error relative to controls by 0.4243 and 0.2090; kernel velocity gives similar reductions of 0.3972 and 0.1974. The discrete coefficient gives the largest point reductions. At 1,024 updates, residual action worsens forecasts on both tasks. This supports a local role in explaining departure, not long-term predictive benefit. Appendix L reports every diagnostic, validation-selected outcomes, and 200 divergent trajectories, all at the largest step scale.

No demonstrated decision advantage. We compare continued nonlinear training with ordinary validation checkpoint selection, kernel freezing, a validation race between the two, and three learned selectors. Each policy receives 30 seconds of online compute and five validation assessments on the same examples. We charge diagnostic and tangent-construction costs to that budget. Across 120 paired cases, all three learned selectors choose continued training every time. Table 7 reports the policy comparisons. The combined selector's mean test-MSE reduction relative to ordinary continuation is -0.00032 (95% interval $[-0.00267, 0.00176]$); relative to the validation race, it is -0.00809 $[-0.02449, 0.00063]$. We average tasks equally and bootstrap datasets and shared seed blocks. Four real-data families give limited uncertainty resolution. Small differences among selectors reflect overhead and checkpoint timing, not successful action selection.

Every policy selects the probe checkpoint in every CIFAR and diabetes case. Sparse later checkpoints may miss short useful training intervals. The diabetes loader also uses full-dataset covariate scaling before split-specific normalization (Appendix I). These limitations narrow the decision test. We therefore report failure of these selectors under this protocol, not an impossibility result for diagnostic control. We did not deploy the uncalibrated signed rule as a policy.

Mechanistic follow-through and computational limits. A separate completed study measures 5,376 cases across eight regression datasets, two activations, two widths, two training sizes, and two initial step scales. At initialization and update 800, we compare nonlinear and matched tangent continuations at 5 and 100 updates. Of 10,752 anchored comparisons, 2,922 have absolute risk gaps above 0.001 at both horizons; 65 change from benefit to harm. An exact endpoint decomposition of the sign margin associated with Equation (17) attributes the largest absolute component to prediction-direction change in 37 reversals and to remaining-error change in 28. This is descriptive accounting, not causal mediation (Appendix K).

The same study tests whether a transported gradient-defect correction can reproduce a prescribed GD trajectory efficiently. On 2,688 cases from four additional task families, it meets the normalized root mean square (RMS) tolerance 0.005 in every case, including 316 charged exact-GD restarts.

Its aggregate cost shows no saving over the reference GD implementation. Faster adaptive Runge–Kutta updates meet the tolerance in only 93.8% of cases. Thus approximation fidelity, predictive quality, and computation cost require separate evaluation; Table 9 reports the comparison.

7 Limitations and conclusion

Residual-action analysis identifies the leading task-visible departure from a frozen kernel under smooth scalar predictions, squared loss, and constant preconditioning. The matched construction removes initialization confounds, while symmetry and discrete curvature identify limits of the leading flow coefficient. The signed contrast and exact risk identity explain why departure and improvement are different outcomes.

The practical contribution is a controlled way to investigate feature learning: match the starting predictor and kernel, measure the residual action, and evaluate held-out benefit with validation-based checkpoint selection. The evidence supports local prediction and mechanistic interpretation within the tested small-model settings. Residual action adds short-horizon information beyond initialization controls, while the discrete expansion identifies the curvature correction needed for actual updates. It does not establish a transferable controller, a robust Adam advantage, or an aggregate speedup in the broader trajectory benchmark. Conditional synthetic gains and negative controls remain part of the evidence. Appendix J gives the expanded optimizer comparisons, including a diabetes kernel-fit result for which ordinary linear ridge performs better.

Reproducibility statement

Appendices A, B, C, D, and E state assumptions and give proofs. Appendices G, I, J, K, and L specify data processing, paired seeds, candidate grids, statistical units, numerical checks, and computational budgets. The anonymous supplementary archive provides experiment and analysis code, frozen designs, and the aggregate and per-seed data used in the displays. The scripts `build_results.py`, `build_followup.py`, `build_mechanisms.py`, and `build_local_dynamics.py` regenerate the corresponding figures and tables. We distinguish fixed primary comparisons from exploratory analyses and retain negative results and numerical amendments.

AI use statement

We used OpenAI Codex (GPT-6) for code development and review, experiment design and execution, mathematical derivations and proof drafting, statistical analysis, and manuscript preparation. Scientific Agent Skills (Kassis et al., 2026) supported evidence tracking and document preparation. We used the Humanizer guidelines to revise the prose. These tools assisted the work and are not authors.

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://arxiv.org/abs/2111.00034.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d027a5c93d484a4312cc486d399c62c1-Abstract-Conference.html.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: A flexible approach to scale hyper-parameters of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/720e7ebc49c84252abc0754dddf80976-Abstract-Conference.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1812.07956.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.13409.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4542–4551, 2020. URL https://proceedings.mlr.press/v119/huang20l.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.07572.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents, 2026. URL https://arxiv.org/abs/2609.00065.

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

Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32617–32648, 2025. URL https://proceedings.mlr.press/v267/lauditi25a.html.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1902.06720.

Philip M. Long. Properties of the after kernel, 2021. URL https://arxiv.org/abs/2105.10585.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), 2018. URL https://arxiv.org/abs/1804.06561.

Noa Rubin, Kirsten Fischer, Javed Lindner, David Dahmen, Inbar Seroussi, Zohar Ringel, Michael Krämer, and Moritz Helias. From kernels to features: A multi-scale adaptive theory of feature learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52225–52257, 2025. URL https://proceedings.mlr.press/v267/rubin25a.html. Author list follows the correction in arXiv:2502.03210.

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.

Scikit-learn contributors. Toy datasets: Diabetes dataset, 2026. URL https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset. Documentation accessed September 21, 2026.

UCI Machine Learning Repository. Wine, 1991. URL https://archive.ics.uci.edu/dataset/109/wine.

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

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

A Complete proofs of the local statements

A.1 Kernel derivative and the quadratic certificate

Since A is constant and symmetric, $K_{ij} = g_i^\top A g_j$ and $\dot{g}_i = H_i \dot{\theta}$. Substituting $\dot{\theta} = -A \sum_k g_k r_k / N$ yields

$$\dot{K}_{ij} = \dot{\theta}^\top H_i A g_j + g_i^\top A H_j \dot{\theta} = -\frac{1}{N} \sum_k r_k (g_k^\top A H_i A g_j + g_i^\top A H_j A g_k).$$

This proves Proposition 1. It is a conditional statement at fixed parameters: changing labels changes the residual and hence the instantaneous vector field.

Differentiate $\dot{f} = -Kr/N$, using $\dot{r} = \dot{f}$:

$$\ddot{f} = -\dot{K} r/N + K^2 r/N^2.$$

The frozen derivative is $\ddot{f}^{\mathrm{K}} = K_0^2 r^{\mathrm{K}}/N^2$. At zero, both residuals and kernels match, so $e_0 = \dot{e}_0 = 0$ and $\ddot{e}_0 = -\dot{K}_0 r_0/N$. Local C^3 smoothness gives continuous third derivatives along the compact trajectory interval and Taylor's integral remainder. Consequently,

$$R_3(t) = \frac{1}{2} \int_0^t (t-s)^2 e_s^{(3)} \, ds, \qquad \|R_3(t)\| \le \frac{t^3}{6} \sup_{0 \le s \le \tau} \left\| e_s^{(3)} \right\|.$$

If $a = \left\| \dot{K}_0 r_0 \right\| > 0$ and $C_\tau > 0$, take $\tau' = \min(\tau, a/(4NC_\tau))$. The reverse triangle inequality gives the lower bound in Theorem 2. If $C_\tau = 0$, any $\tau' \le \tau$ suffices.

An explicit finite remainder bound. Let $B_j = \sup_{s \in [0,\tau]} \left\| K_s^{(j)} \right\|_{\mathrm{op}}$ for $j = 0, 1, 2$, and $R = \|r_0\|$. Along the nonlinear flow,

$$\frac{d}{dt} \|r\|^2 = -2 r^\top K r/N \le 0,$$

and similarly for the frozen flow. Differentiating again gives

$$f^{(3)} = -K'' r/N + (2K'K + KK')r/N^2 - K^3 r/N^3. \tag{18}$$

For the frozen model, $(f^{\mathrm{K}})^{(3)} = -K_0^3 r^{\mathrm{K}}/N^3$. One valid choice is therefore

$$C_\tau = \frac{R}{6} \left[\frac{B_2}{N} + \frac{3 B_0 B_1}{N^2} + \frac{B_0^3 + \|K_0\|_{\mathrm{op}}^3}{N^3} \right]. \tag{19}$$

The constants depend on kernel derivatives along the trajectory. This bound is conditional on those controls and does not provide an initialization-only guarantee of a useful forecast horizon.

A.2 Finite-time control and the third derivative

Subtracting the two output flows yields

$$\dot{e}_t = -K_t e_t/N - (K_t - K_0) r_t^{\mathrm{K}}/N.$$

Variation of constants proves Equation (4). For $z(t) = U(t,s) z_s$, $d\|z(t)\|^2/dt = -2 z(t)^\top K_t z(t)/N \le 0$, so $\|U(t,s)\|_{\mathrm{op}} \le 1$. Also $\left\| r_s^{\mathrm{K}} \right\| \le \|r_0\|$. These facts prove Equation (5), without assuming commutation. In particular, $\|K_s - K_0\|_{\mathrm{op}} \le Ls$ implies $\|e_t\| \le L \|r_0\| t^2/(2N)$.

At time zero, subtract the frozen third derivative from Equation (18); the $K_0^3 r_0$ terms cancel. The two remaining mixed products have coefficients two and one, respectively, giving Equation (6). If $\dot{K}_0 r_0 = 0$, the term $K_0 \dot{K}_0 r_0$ vanishes, but $\dot{K}_0 K_0 r_0$ need not vanish. Taylor expansion with continuous third derivative gives the stated $o(t^3)$ remainder.

A.3 Stationarity and clock invariance

If $K_0 r_0 = 0$, then $0 = r_0^\top K_0 r_0 = \left\| A^{1/2} J_0^\top r_0 \right\|^2$. Hence $A J_0^\top r_0 = 0$ and the initial vector field vanishes. The locally Lipschitz autonomous vector field admits the constant solution uniquely. At a nonstationary initialization $K_0 r_0 \neq 0$, both ν_{eff} and χ are defined. Replacing A by cA multiplies the vector field by c and the kernel by c. This rescaling multiplies its directional derivative along the vector field by c^2. Substitution proves the clock invariance of χ. The displacement ratio follows by dividing $t^2 \left\| \dot{K}_0 r_0 \right\| / (2N)$ by $t \left\| K_0 r_0 \right\| / N$.

B Matched particles and symmetry

B.1 Explicit velocity average and width law

The centering in Equation (8) gives zero initial predictions for every exponent. A readout derivative is $m^{-\alpha}\phi(w_a^\top x)$, and a hidden-weight derivative is $m^{-\alpha} c_a \phi'(w_a^\top x) x$. Their Gram matrix multiplied by $m^{2\alpha-1}$ is Equation (9). This equality is at a common parameter value; later parameters generally differ between exponents.

At initialization $r_0 = -y$. Set $s_{ij} = x_i^\top x_j$ and

$$b_a = \frac{1}{N} \sum_k y_k \phi_{ak}, \qquad v_a = \frac{c_a}{N} \sum_k y_k \phi'_{ak} x_k.$$

The particle velocities are $\dot{c}_a = m^{\alpha-1} b_a$ and $\dot{w}_a = m^{\alpha-1} v_a$. Differentiating the common kernel gives $V_{m,ij} = m^{-1} \sum_a D_{a,ij}$, where

$$\begin{aligned} D_{a,ij} = {} & 2 c_a b_a \phi'_{ai} \phi'_{aj} s_{ij} \\ & + (x_i^\top v_a)(\phi'_{ai}\phi_{aj} + c_a^2 \phi''_{ai}\phi'_{aj} s_{ij}) \\ & + (x_j^\top v_a)(\phi_{ai}\phi'_{aj} + c_a^2 \phi'_{ai}\phi''_{aj} s_{ij}). \end{aligned} \tag{20}$$

This proves the exact finite-width law. For fixed data, each summand is a function of one initial particle. Under the stated first-moment conditions the strong law applies entrywise; fixed finite N converts entrywise convergence to Frobenius convergence. Nonzero limiting numerator and denominator give the almost-sure Θ claim. Applying the matrices to fixed r_0 proves the task-effective version when its limits are nonzero. In particular,

$$\frac{d^2}{dt^2}(f_t^{(\alpha)} - f_t^{\mathrm{K}})\Big|_{t=0} = -m^{\alpha-1} V_m r_0 / N.$$

Under nondegeneracy, the leading acceleration vanishes as $m^{-1/2}$ at $\alpha = 1/2$ and stays order one at $\alpha = 1$.

Ordinary non-centered networks. For $\widetilde{f}_{m,\alpha}(x) = m^{-\alpha} \sum_a c_a \phi(w_a^\top x)$, the derivative and effective-kernel calculations above remain unchanged. However, $\widetilde{r}_{\alpha,0} = m^{-\alpha} \sum_a c_{a,0} \phi(w_{a,0}^\top X) - y$ now depends on α. Replacing y by $-\widetilde{r}_{\alpha,0}$ in the particle responses gives $\dot{K}_0^{(\alpha)} = m^{\alpha-1} V_m(\widetilde{r}_{\alpha,0})$, where $V_m(\cdot)$ is a common linear map of the residual. The prefactor persists, but the full cross-exponent equality needs matched residuals. Centering supplies that match without changing derivatives. Matching residuals by changing labels is an algebraic control, not a generalization comparison on a common task.

B.2 Population and exact cancellation

Under $c \mapsto -c$, b_a is unchanged and v_a changes sign. Every term in Equation (20) changes sign. If ϕ is odd, then under $w \mapsto -w$, ϕ and ϕ'' change sign, ϕ' does not, b_a changes sign, and v_a does not. Again D_a is odd. In either case its mean is zero. Finite second moments give $V_m = O_p(m^{-1/2})$ entrywise and hence in norm. Dividing by the nonvanishing kernel norm proves $\nu_K = O_p(m^{\alpha-3/2})$. The task-effective statement follows identically with a nonzero limiting denominator.

For exact pairs, cancellation holds for every residual vector, not just $-y$. Let $s = m^{\alpha-1}$ and express the particle vector field as $\dot{\theta} = s\, b(\theta, r)$; at initialization b is independent of α. The common kernel is a function of parameters alone. At zero,

$$K'' = s^2\{D^2K[b,b] + DK[D_\theta b\, b]\} + s\, DK[D_r b\, \dot{r}].$$

The last term is zero: $D_r b\, \dot{r}$ is the same particle-response construction driven by another residual, and its kernel derivative cancels pairwise. The expression in braces defines W_m, independent of α. Since $K_0' = 0$, Equation (6) reduces to $e_0^{(3)} = -s^2 W_m r_0/N$. This proves Proposition 5. We require an additional assumption to assert a nonzero limit of W_m.

C Exact rank-one closure and alignment

For $L(a, w) = \|aw - u\|^2/2$ and $q = aw$, gradient flow is

$$\dot{a} = -w^\top(q - u), \qquad \dot{w} = -a(q - u).$$

Consequently $d(a^2 - \|w\|^2)/dt = 0$. Since $\|q\|^2 = a^2(a^2 - \delta)$, solving for the nonnegative a^2 gives $\gamma(q) = (\delta + \sqrt{\delta^2 + 4\|q\|^2})/2$. Wherever $\gamma > 0$, $ww^\top = qq^\top/\gamma$ and

$$\dot{q} = \dot{a}w + a\dot{w} = -(ww^\top + a^2 I)(q - u).$$

This is Equation (12). At $q = 0$ with $\delta > 0$, $w = 0$ and $a^2 = \delta$, so the same closure holds directly. The input-space tangent kernel is

$$k_t(x, x') = x^\top(a^2 I + ww^\top)x'.$$

Its eigenvalue along q is $\gamma + \|q\|^2/\gamma$, with eigenvalue γ in perpendicular directions and eigengap $\|w\|^2$.

For $q \neq 0$, let $\widehat{q} = q/\|q\|$ and $\widehat{u} = u/\|u\|$. Differentiating $\rho = \widehat{q}^\top \widehat{u}$ gives

$$\dot{\rho} = \frac{1}{\|q\|}\widehat{u}^\top(I - \widehat{q}\widehat{q}^\top)\dot{q}.$$

The rank-one term and the component proportional to q vanish under projection, leaving $\gamma\|u\|(1 - \rho^2)/\|q\|$. Thus signed cosine increases for $|\rho| < 1$. Absolute cosine increases when $\rho > 0$ and decreases when $\rho < 0$; the kernel axis is unoriented.

When $\delta > 0$, $\gamma \geq \delta$, and

$$\dot{L} = -(q - u)^\top(\gamma I + qq^\top/\gamma)(q - u) \leq -2\delta L.$$

Gronwall's inequality proves exponential loss decay. Decreasing loss bounds q; moreover $a^2 + \|w\|^2 = \sqrt{\delta^2 + 4\|q\|^2}$, so parameters stay bounded. The smooth parameter ODE therefore extends globally, justifying the convergence statement rather than only a conditional interval bound.

D Proof of the held-out risk expansion

Let $r_X = f_X - y_X$ and $r_V = f_V - y_V$. The held-out output obeys $\dot{f}_V = -B_t r_X/N$. The frozen comparator satisfies $\dot{f}_V^{\mathrm{K}} = -B_0 r_X^{\mathrm{K}}/N$, where $\dot{r}_X^{\mathrm{K}} = -K_0 r_X^{\mathrm{K}}/N$. At zero the outputs and first derivatives match. Differentiating,

$$\ddot{f}_V = -\dot{B}_t r_X/N + B_t K_t r_X/N^2, \qquad \ddot{f}_V^{\mathrm{K}} = B_0 K_0 r_X^{\mathrm{K}}/N^2.$$

Their difference at zero is $-\dot{B}_0 r_{X,0}/N$, proving Equation (14) by the same Taylor argument. With $e_V = f_V - f_V^{\mathrm{K}}$,

$$R_V - R_V^{\mathrm{K}} = \{2(r_V^{\mathrm{K}})^\top e_V + \|e_V\|^2\}/N_V.$$

Use $r_V^{\mathrm{K}} = r_{V,0} + O(t)$ and $e_V = -t^2\dot{B}_0 r_{X,0}/(2N) + O(t^3)$ to obtain Equation (15). A nonzero coefficient dominates its remainder on a sufficiently short interval. The theorem gives no sign if the coefficient vanishes. The coefficient involves held-out residuals and therefore requires held-out labels; it is not a theorem about an unknown population distribution.

E A FINITE-STEP COUNTERPART OF THE LOCAL CERTIFICATE

Assume C^3 predictions near initialization and a constant symmetric positive semidefinite A. Compare k full-batch GD updates on $L = \|f_X - y_X\|^2/(2N)$ with GD on the affine model anchored at the same θ_0, using the same A and step η. Fix k as $\eta \to 0$. Let

$$g_0 = J_0^\top r_{X,0}/N, \qquad u_0 = Ag_0, \qquad H_L = \nabla^2 L(\theta_0), \qquad G_0 = J_0^\top J_0/N.$$

Taylor expansion of the parameter updates gives

$$\theta_k - \theta_0 = -k\eta u_0 + \binom{k}{2}\eta^2 A H_L u_0 + O(\eta^3), \qquad \delta_k = -k\eta u_0 + \binom{k}{2}\eta^2 A G_0 u_0 + O(\eta^3),$$

where δ_k is the affine model's parameter increment. These formulas follow by induction: expand each gradient at θ_0 and sum the preceding first-order increments. Expanding the held-out nonlinear predictor and subtracting $f_{V,0} + J_{V,0}\delta_k$ yields

$$e_{V,k} = \binom{k}{2}\eta^2 J_{V,0} A(H_L - G_0)u_0 + \frac{k^2\eta^2}{2} h_{V,0} + O(\eta^3).$$

The residual part of the loss Hessian satisfies $H_L - G_0 = N^{-1}\sum_i r_{X,0,i}\nabla^2 f(x_i;\theta_0)$. Since the flow direction is $-u_0$, differentiating $B = J_V A J_X^\top$ gives

$$-\dot{B}_0 r_{X,0}/N = h_{V,0} + J_{V,0} A(H_L - G_0)u_0.$$

Substitution proves Equation (16). At $k = 1$, the accumulated residual-Hessian term vanishes, leaving the output-curvature contribution. The fixed-k remainder makes no claim about horizons that grow as $1/\eta$.

Independent analytic and autodiff checks compare the Jacobian, its directional derivative, the loss gradient, and the residual-Hessian action for tanh and shifted-softplus two-layer networks, with both centered and non-centered outputs and a fixed layerwise metric. At $k = 1, 2, 8$, step halving gives the expected cubic error reduction. These checks test the local expansion; the separate longer-horizon experiments do not follow from its remainder bound.

F IMPLEMENTATION AND NUMERICAL VALIDATION

Action-only differentiation. For a parameter vector θ and fixed residual r, evaluate $J^\top r$ by a vector–Jacobian product, multiply by diagonal A, and evaluate $J(AJ^\top r)$ by a Jacobian–vector product. Differentiate this composite action along $v = -AJ^\top r/N$, with r detached. For the held-out action, use the training VJP and held-out JVP; the vector field still uses training data only. The implementation in `improved/core.py` never materializes a Jacobian or kernel in these routines. Evaluating the geometric Frobenius diagnostic, or exact kernel regression, is a different task and may require dense matrices.

For the third-order gap, define the vector field $v(\theta)$ and recursively $F_1 = Df[v]$, $F_2 = DF_1[v]$, $F_3 = DF_2[v]$. Subtract $(-K_0/N)^j r_0$ for $j = 2, 3$. Differentiating the full composite functions includes the derivatives of v, which are needed in Equation (6).

Exact continuation and stable arithmetic. At refresh time s, let $K_s = Q\Lambda Q^\top$ be the training kernel, B_s the evaluation-to-training kernel, r_s the current training residual, and b_s the current evaluation output. At continuation time h,

$$f_{s,h}^{\text{refresh}} = b_s - B_s Q \operatorname{diag}\left(\frac{1 - e^{-\lambda_j h/N}}{\lambda_j}\right) Q^\top r_s, \tag{21}$$

where the zero-eigenvalue limit is h/N. We use `-expm1` for small arguments, symmetrize the numerical training Gram matrix, and clip roundoff-negative eigenvalues to zero. For centered initialization $f_0 = 0$, initial-kernel ridge predictions are $B_0(K_0 + N\lambda I)^\dagger y$ with the numerical pseudoinverse convention in the code. These details matter when a nearly converged flow has an almost flat validation or test-error curve.

Checks and benchmark. The 1,536 independent action checks use $N = 64, d = 16$, widths $32, 64, \ldots, 4096$, $\alpha \in \{0.5, 0.75, 1\}$, tanh/shifted softplus, and standard/exact-pair initializations, over 16 seeds. We differentiate each exponent independently and never infer one action from another exponent. Explicit-kernel directional differentiation and centered finite differences provide independent comparisons. The largest absolute matrix-free versus explicit action discrepancy is 9.18×10^{-16}. Five implementation tests cover action correctness, clock rescaling, stable continuation, cubic derivatives, and the held-out formula including nonzero initial predictions; the nine reference tests cover the analytic identities.

The separate cubic study has four seeds, three widths, three exponents, and three times (108 observations). At $t = 0.01$, the mean norm ratio is 0.9979318, the minimum cosine is 0.99999984, and the largest computed second-order gap derivative is 3.20×10^{-17}. Four paired primary seeds integrated with step 0.02 and 0.01 differ by at most 3.60×10^{-10} in selected adaptive test MSE and 9.46×10^{-10} in refreshed test MSE.

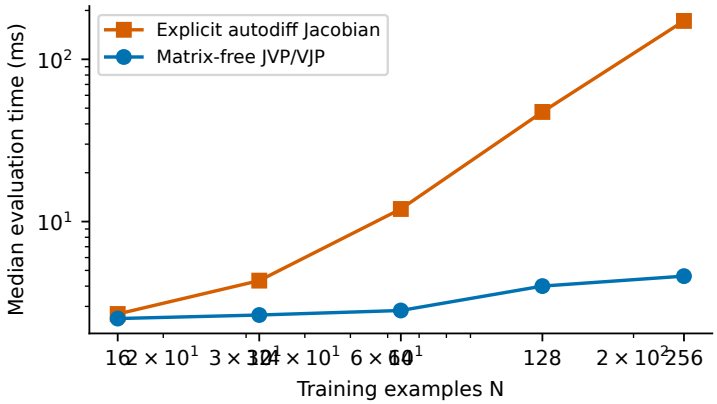


Figure 3: CPU certificate benchmark: $d = 16$, width 256, four threads, one warmup and five timed repetitions in a fresh process for each method and size. Curves show medians, not uncertainty intervals. At $N = 256$, explicit autodiff takes 172.43 ms versus 4.61 ms matrix-free. The comparison is against explicit autodiff, not an optimized analytic formula.

Figure 3 compares evaluation time across sample sizes. At $N = 256$, whole-process peak resident set size is approximately 1.58 GiB for explicit differentiation and 0.65 GiB for matrix-free evaluation. These values include interpreter and framework memory and must not be read as isolated tensor allocation. Benchmark timings are hardware-specific and do not establish end-to-end predictive training savings.

G Experimental protocols and evidence provenance

G.1 Reference synthetic protocols

We preserve the reference analytic implementation and independent autodiff checks separately. Agreement between implementations tests the formulas, while the paired prediction studies assess their interpretation. The supplementary archive records per-seed measurements, source versions, and the numerical checks summarized below.

The reference experiments use float64 full-batch flow, mainly explicit midpoint integration, with RK4 for local robustness checks. The drivers evaluate frozen-kernel training and cross predictions by eigendecomposition. The Gaussian teacher is $x \sim \mathcal{N}(0, I_d/d)$, $y = \tanh(gu^\top x)$, $\|u\| = \sqrt{d}$.

The width experiment uses $N = 64, d = 16$, widths 32 through 4096 in powers of two, ten paired seeds, shifted softplus $\phi(z) = \log(1 + e^z) - \log 2$, $c_0 = 1$, and Gaussian hidden weights. The driver truncates a maximal-width initialization to give nested widths and measures drift at time 3 with step 0.02. The Taylor study uses $N = 48, d = 12, m = 512$, six seeds, and times 0.005 through 0.12. The generalization study uses tanh, $N = 128, d = 50$, 2048 independent test points, widths 256

through 2048, eight seeds, and time 50 with step 0.02. The covariance statistic is $C_t = W_t^\top W_t/m$; feature spike is $\lambda_{\max}(C_t - C_0)$ and alignment is the absolute cosine of its leading eigenvector with u. The rotation intervention changes labels at fixed inputs/initialization over seven teacher angles, with five paired seeds and time 30. Permutation retains inputs and label values while disrupting their association.

The reference shifted-softplus helper uses PyTorch's default linear threshold at 20 while its analytic derivative helper uses sigmoid everywhere. Thus the implementation is not exactly the ideal globally C^3 function at the threshold. This did not prevent numerical reproduction at the sampled scales, but the mathematical statements assume local smoothness and should not be interpreted as a global guarantee for thresholded implementations.

All principal synthetic MSE values reproduce the recorded CSV precision. One frozen-oracle argmin time differs by about 5,425 on a numerically flat converged tail, with identical selected test MSE. Taylor ratios differ by at most 6.44×10^{-8}. The symmetry table reproduces exactly; antithetic cubic norm ratios differ by at most 8.18×10^{-5} because extremely small gaps involve subtraction, while the headline cubic behavior is unchanged. Deep-network ratios differ by at most 3.12×10^{-8}. These are numerical, not bitwise environment, reproductions.

G.2 Validation-controlled prediction study

For each seed, independent pseudorandom streams use base $10^6 + \text{seed}$ with offsets $10^5, 2 \cdot 10^5, 3 \cdot 10^5, 4 \cdot 10^5, 5 \cdot 10^5, 6 \cdot 10^5$ for teacher, train inputs, validation inputs, test inputs, training noise, and weights. Teacher normalization fixes $\|u\| = \sqrt{d}$. Methods share all data and initialization within a seed; gains share the same underlying draws. Main settings are $N = 128, d = 50, m = 2048$, 256 validation and 2048 test examples, tanh, $c_0 = 1$, and $w_0 \sim \mathcal{N}(0, I_d)$. We center predictions on every evaluated input set.

Adaptive training uses explicit midpoint step 0.02 to time 200. Candidate checkpoints are $5, 10, 25, 50, 100, 200$. The frozen continuation grid is the union of $\{0, 50\}$ and 401 logarithmically spaced times from 10^{-2} to 10^6. Ridge penalties are zero and 61 logarithmically spaced values from 10^{-9} to 10^1, in the $K/N + \lambda I$ convention. Refreshed kernels use checkpoints $5, 10, 25, 50$ and the same continuation grid, applying Equation (21). Each method selects by validation MSE only, with first-grid-entry tie breaking. We use test predictions solely for final scoring and separately labeled common-clock diagnostics.

The primary study has seeds 0–63 at each gain. The precision reserve begins at seed 64, uses the same fixed design, and stops at the wall-clock budget. Seeds 64–618 complete all four gains, giving 555 balanced seeds per gain. Seed 619 completes gains 2 and 3 only; we retain those two rows with `analysis_eligible=false`. We report primary and precision results separately. Two preliminary pilot seeds (9000 and 9001) informed feasibility work; we exclude them from all primary and replication summaries.

The extended-horizon study reuses primary seeds 0–15, adding checkpoints 500 and 1,000. Because many selected the upper boundary, an exploratory audit reuses seeds 0–3 and adds 2,000, 5,000 and 10,000. Early refresh candidates and frozen/ridge grids are unchanged. At horizon 1,000, validation selection chooses the boundary in 14/16, 10/16, 8/16 and 7/16 cases for gains 2–5. The four-seed audit is not an estimate of uniformly converged behavior: some selected checkpoints fit training labels to numerical precision, while others select earlier checkpoints.

Eight secondary cells each use 16 paired seeds at gains 2 and 4: $N = 64$, $N = 256$, additive Gaussian training noise of standard deviation 0.1 or 0.3, shifted softplus, exact antithetic pairs, width 512, and width 4096. All other defaults remain fixed. In the antithetic cell, we duplicate hidden weights in pairs and negate half the readouts. In noisy cells, validation and test labels remain clean. These cells probe sensitivity rather than independent confirmatory hypotheses.

G.3 Statistical analysis

The unit of paired inference in synthetic prediction is a seed. We report mean test MSE and mean paired differences. Percentile bootstrap intervals resample paired seeds 2,000 times; a deterministic random stream keyed by study, gain, and comparison prevents unrelated queue growth from chang-

ing an already completed study's interval. Intervals quantify variation over the specified synthetic data/teacher/initialization generation, not arbitrary tasks. One-sided Wilcoxon signed-rank tests target positive improvement, with Holm correction across gains within each study and comparison. The fixed primary endpoint is refreshed versus adaptive at horizon 200. Other contrasts, secondary conditions, and horizon audits are exploratory; we do not apply a global correction across exploratory cells. Four-seed positive bootstrap intervals should not be interpreted as confirmatory significance: the minimum exact one-sided signed-rank p with four nonzero concordant pairs is 0.0625.

G.4 Held-out signed-risk checks

This exploratory identity study uses 16 seeds, $N = 32$, $N_V = 128$, $d = 12$, width 128, $\alpha = 1$, and tanh/shifted softplus. The held-out labels always follow the gain-2 teacher. Training labels are either teacher labels, their random permutation, or a Gaussian vector rescaled to the teacher training-label norm. The three label conditions and two activations share data within a seed. Local trajectories use RK4 step 10^{-4} at times $0.005, 0.01, 0.02, 0.05$ (384 observations); the extended trajectories use step 0.005 at $0.1, 0.5, 1, 5, 10, 50$ (576 observations). We evaluate the signed coefficient only at initialization and score a sign as resolved if its predicted absolute MSE contrast exceeds 10^{-11}; all local observations satisfy this condition. The mean absolute contrast error is approximately 3.05×10^{-11} at 0.005 and 3.03×10^{-8} at 0.05, consistent with the expected cubic local error. We use this study to check the identity, without selecting a predictive model.

G.5 CIFAR-10 scope check

The CIFAR drivers use five class pairs: airplane–automobile, bird–cat, deer–dog, frog–horse, and ship–truck, with three seeds each. We use CIFAR-10 (Krizhevsky, 2009) and verified the extracted archive against MD5 `c58f30108f718f92721af3b95e74349a`. The local certificate uses 48 training images per class, convolutional channels (4,8), two 5×5 convolutions with softplus and average pooling, followed by a scalar readout. We center predictions at initialization. RK4 step 0.0005 evaluates times $0.0025, 0.005, 0.01, 0.02, 0.04$.

The performance driver uses 128 training, 128 validation, and 1000 test images per class, and nonlinear checkpoints $50, 100, 200, 500, 1000$ with midpoint step 0.25. We preserved the reference class-pair indexing and seeds when splitting the driver into independent jobs. Chunked Jacobian evaluation reduces memory use without changing the computed model. Accuracy counts match the reference measurements, test MSE differences are about 10^{-12} or less, and local norm-ratio discrepancies are at most 1.13×10^{-9}.

For the validation-selected MSE comparison, we first average seed effects within task and then bootstrap-resample the five tasks. Task-level frozen-minus-adaptive differences are -0.000962, 0.005102, 0.004741, 0.018918, and -0.009077 in the pair order above. Their mean is 0.003744 with 95% task-bootstrap interval $[-0.003550, 0.012107]$. Five fixed tasks sharing a dataset do not establish population-wide independence or broad image-classification improvement. These experiments compare the reference nonlinear and initial-kernel methods only.

G.6 Execution scope and resource limits

The scheduled run lasted 28,801.6 seconds (approximately eight hours), from 18:28 UTC to 02:28 UTC. A single NVIDIA RTX 5090 with 32 GiB VRAM ran the main queue; auxiliary numerical studies used CPU resources. The environment was Python 3.12.3, PyTorch 2.12.1+cu130, NumPy 2.5.3, SciPy 1.18.1, pandas 3.0.6, and matplotlib 3.11.2 on an Intel i9-12900K host. We report numerical agreement rather than bitwise equivalence across software environments.

The supervisor completed 2,787 queued jobs, found one existing completed output, and left 5,714 optional reserve jobs unstarted. The result collection contains 2,822 completed extension records because some queue jobs also execute paired horizon audits. All mandatory reproduction, primary, secondary, and horizon stages completed. We exclude unstarted reserve jobs and partial checkpoints from the evidence. At approximately 00:37 UTC the shared filesystem filled; this study occupied only about 1.3 GiB of the 271 GiB filesystem. We recovered from checkpoint failures and a supervisor exit by removing redundant downloaded archives, compacting JSON whitespace without

changing values, and resuming from the preserved ledger. We reran every affected case and resolved all failures before the reference cutoff. The execution ledger records this brief interruption.

H SYNTHETIC MECHANISMS AND EXTENDED PREDICTION RESULTS

Table 3 reports the matched-regime measurements, and Figure 4 shows the width, time, and rank-one checks. Figure 5 tests task alignment through label interventions. Table 4 separates the common-time advantage from stopping-time effects. Tables 5 and 6 report horizon and sensitivity analyses. Figure 6 examines symmetry, depth, and local diagnostic correlations; its pooled correlations do not control for shared width and scaling effects.

α	Velocity slope [95% CI]	Drift slope	Test MSE	Spike	Alignment
0.50	-0.511 $[-0.532, -0.494]$	-0.507	0.29372	0.00976	0.68206
0.75	-0.261 $[-0.280, -0.244]$	-0.259	0.28263	0.21770	0.85591
1.00	-0.011 $[-0.030, 0.006]$	0.0147	0.16044	6.11227	0.93123

Table 3: Matched-regime summary. Slopes use ten paired seeds and widths 32–4096. Prediction and feature statistics use eight seeds, width 2048, and time 50. The common frozen MSE is 0.29393. Weak lazy covariance growth makes its axis alignment poorly identified.

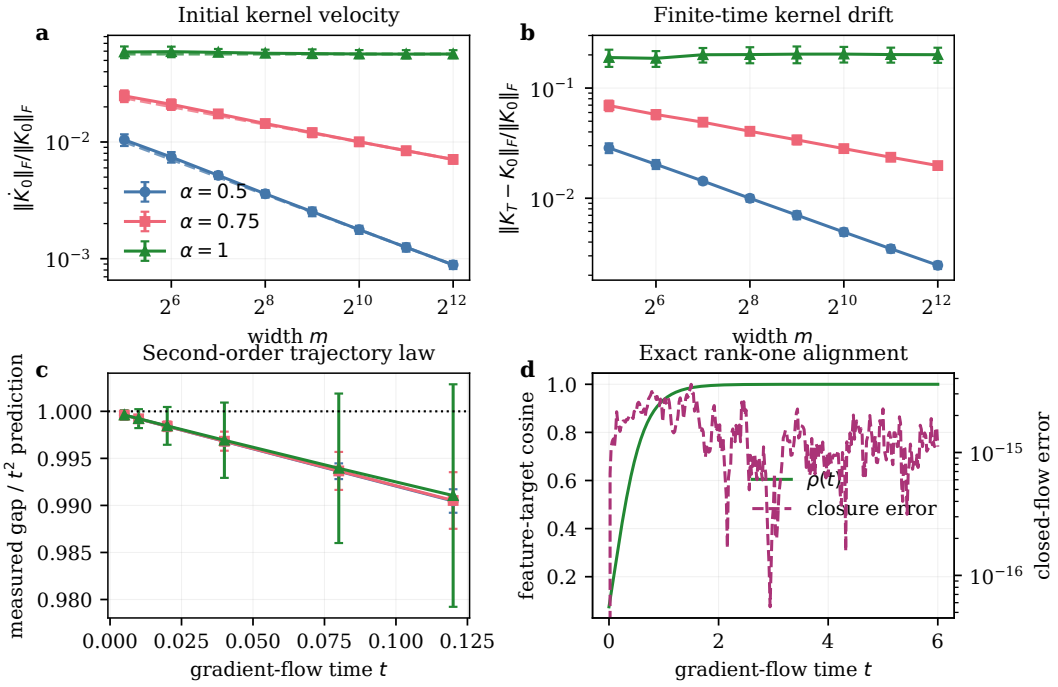


Figure 4: Synthetic theory checks: width velocity, time-3 drift, quadratic trajectory gap, and rank-one alignment/closure. Intervals summarize variation across paired seeds. Different-exponent velocities in this driver use the exact rescaling law; the independent derivative sweep addresses that evaluation limitation.

The smallest-time Taylor ratio is 0.9996048, with minimum direction cosine 0.99999927. Rank-one direct and closed trajectories agree within 3.59×10^{-15}, and imbalance drift is at most 2.47×10^{-14}. Within-seed Spearman correlation between initial ν_{eff} and finite-time linearization gap averages 0.9828 with a 95% t-interval $[0.9731, 0.9924]$ over ten seeds. This is a trajectory-departure correlation, not a held-out-performance prediction. In the label intervention, adaptive teacher-direction covariance growth is 5.2901 for teacher labels and 0.12265 after permutation; the reference Holm-adjusted signed-rank value is 0.0078. Mean adaptive spike-angle error under label rotation is 1.0435 degrees.

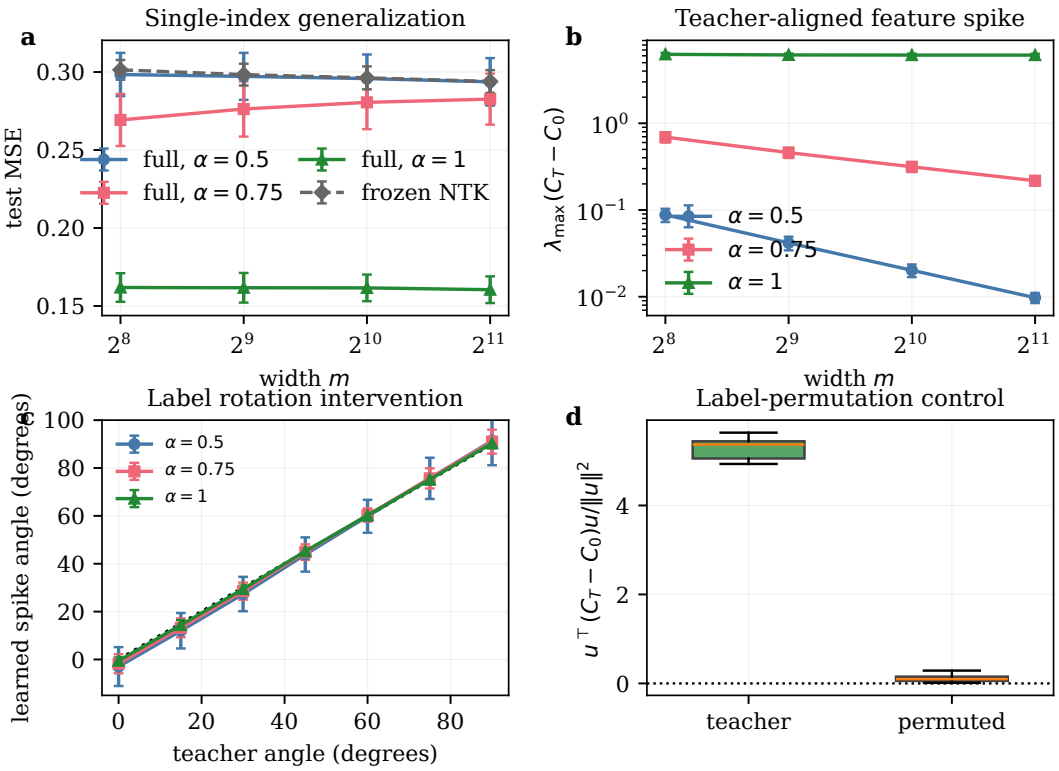


Figure 5: Task-alignment experiments. The adaptive tanh network develops a teacher-aligned covariance spike; label rotation rotates the spike, and label permutation suppresses teacher-direction growth. Panel (c) labels each scaling exponent; its dotted diagonal marks exact tracking of the teacher angle. These are long-time representation results in a population first-velocity-cancelled setting, not a validation of a nonzero initial-velocity predictor.

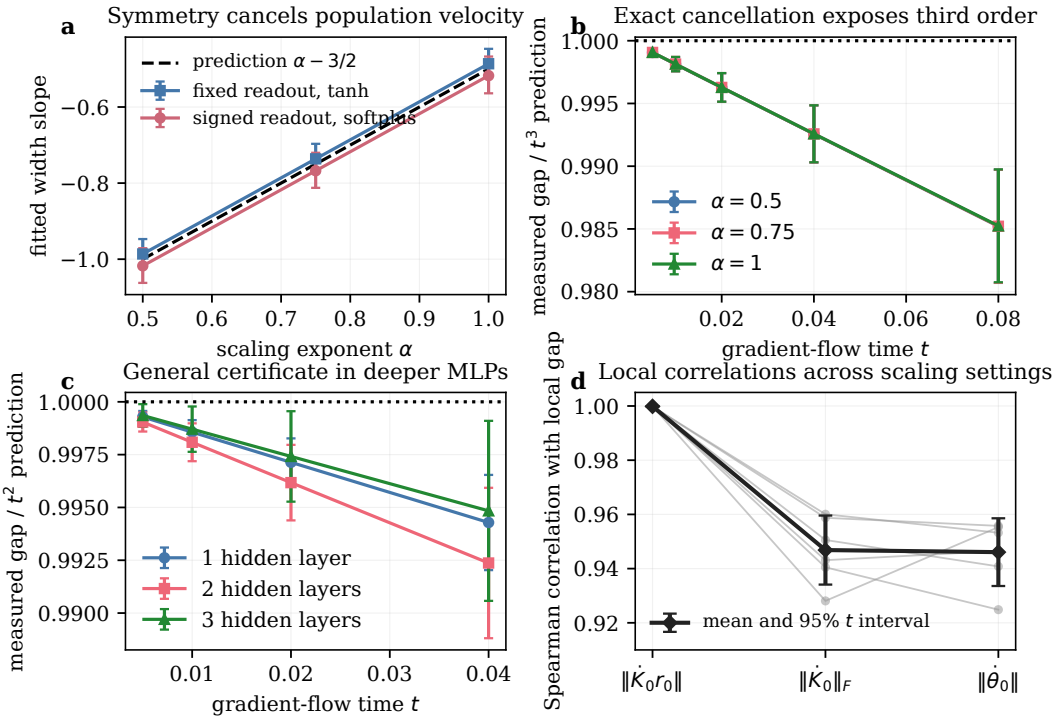


Figure 6: Robustness panels covering symmetry, cubic fallback, depth, and diagnostic specificity. The action-only cubic and held-out studies are separate experiments. Panel (d) pools settings that vary width and scaling exponent within seeds; shared scaling trends can explain high correlations, so these values do not establish incremental predictive utility.

Teacher gain	Adaptive at 50	Frozen at 50	Train-MSE matched	Frozen test oracle
2	0.16044	0.29393	0.16561	0.15564
3	0.23078	0.37540	0.24456	0.23927
4	0.27989	0.42551	0.29864	0.29536
5	0.31473	0.45875	0.33643	0.33407

Table 4: Stopping-time controls, eight paired seeds, width 2048. Entries are mean test MSE. The oracle uses test labels and is not a valid model-selection procedure; we retain it only to expose acceleration effects.

Study / horizon	Gain	Seeds	Adaptive	Refreshed	Improvement [95% CI]
Independent / 200	2	555	0.11540	0.11128	0.00412 [0.00399, 0.00425]
Independent / 200	3	555	0.19014	0.18455	0.00559 [0.00537, 0.00584]
Independent / 200	4	555	0.24430	0.23841	0.00589 [0.00562, 0.00615]
Independent / 200	5	555	0.28319	0.27743	0.00576 [0.00547, 0.00606]
Extended / 1,000	2	16	0.10815	0.10754	0.00061 [0.00005, 0.00115]
Extended / 1,000	3	16	0.18056	0.17702	0.00354 [0.00280, 0.00429]
Extended / 1,000	4	16	0.23291	0.22807	0.00484 [0.00367, 0.00593]
Extended / 1,000	5	16	0.27053	0.26495	0.00558 [0.00407, 0.00717]
Audit / 10,000	2	4	0.10446	0.10380	0.00066 [-0.00027, 0.00193]
Audit / 10,000	3	4	0.17242	0.16864	0.00378 [0.00253, 0.00531]
Audit / 10,000	4	4	0.22077	0.21544	0.00533 [0.00386, 0.00677]
Audit / 10,000	5	4	0.25512	0.24941	0.00570 [0.00372, 0.00769]

Table 5: Independent replication and longer-horizon audits. Improvement is adaptive minus refreshed MSE. The precision expansion uses fresh seeds; the other rows reuse subsets of primary seeds. Four-seed intervals are descriptive and do not establish significance.

Condition	Gain	Frozen	Adaptive	Refreshed	Improvement [95% CI]
n64	2	0.24411	0.22935	0.22758	0.00177 [0.00081, 0.00288]
n64	4	0.41708	0.39858	0.39569	0.00289 [0.00162, 0.00433]
n256	2	0.12412	0.06827	0.05962	0.00866 [0.00794, 0.00942]
n256	4	0.23905	0.14917	0.13826	0.01091 [0.00935, 0.01263]
noise01	2	0.15979	0.12246	0.11795	0.00452 [0.00394, 0.00504]
noise01	4	0.29581	0.25134	0.24439	0.00695 [0.00560, 0.00836]
noise03	2	0.19633	0.17110	0.16712	0.00398 [0.00272, 0.00545]
noise03	4	0.32832	0.29534	0.29204	0.00330 [0.00083, 0.00595]
softplus	2	0.16097	0.17930	0.16303	0.01627 [0.01276, 0.02027]
softplus	4	0.29875	0.33011	0.30183	0.02828 [0.02426, 0.03282]
antithetic	2	0.15457	0.11490	0.11050	0.00439 [0.00368, 0.00506]
antithetic	4	0.29112	0.24234	0.23604	0.00630 [0.00480, 0.00762]
width512	2	0.15591	0.11655	0.11256	0.00399 [0.00333, 0.00464]
width512	4	0.29331	0.24515	0.23952	0.00563 [0.00434, 0.00692]
width4096	2	0.15318	0.11439	0.10988	0.00451 [0.00377, 0.00521]
width4096	4	0.28879	0.24151	0.23505	0.00647 [0.00500, 0.00780]

Table 6: Exploratory secondary studies, 16 paired seeds per condition and gain. Improvement compares refreshed to adaptive. In the softplus rows the frozen initial kernel remains better than both, despite a positive refreshed-over-adaptive contrast. We add noise to training labels only.

I TRANSFER AND TRAINING-DECISION STUDY

I.1 DESIGN, DATA, AND CALIBRATION

We designed this follow-up to test whether diagnostics predict useful adaptation beyond the local Taylor regime and whether they improve a training decision. Calibration uses 12 configurations from synthetic linear/tanh teachers, handwritten digits, and breast-cancer classification, with eight seeds. The calibration set contains 96 benchmark cases and 192 timed actions. We fit the rules to validation outcomes and froze them at 19:48:54 UTC on September 20, 2026, before the first held-out job. We never refit them on held-out results.

The held-out benchmark spans 32 configurations: Fashion-MNIST (Xiao et al., 2017) class pairs 0/6 and 2/4, CIFAR-10 (Krizhevsky, 2009) pairs 0/1 and 3/5, wine (UCI Machine Learning Repository, 1991) pairs 0/1 and 1/2, diabetes regression (Scikit-learn contributors, 2026), and synthetic absolute-value and sine teachers. MLPs use two or three hidden layers of width 48; the small CNN uses channels 4 and 8. The held-out models use tanh. We cap splits at 128 training, 48 validation, and 256 test examples, with smaller stratified splits for tabular classification. We fit the study's additional tabular centering, scale factors, and regression target scaling on training data. However, the default diabetes loader already scales covariates using the full dataset; the later standard-deviation floor prevents exact cancellation of that preprocessing. Diabetes therefore does not provide a strictly inductive preprocessing test. Binary labels take values ± 1, and all tasks use squared loss. We downsample Fashion-MNIST images to 14×14 and CIFAR images to 16×16; their official test splits remain separate from training and validation. These are small-subset experiments, not standard full-dataset accuracy benchmarks.

At update 25, we measure relative Euclidean parameter movement, relative Frobenius change in the hidden-representation Gram matrix, relative kernel drift, χ, and the normalized signed cross-kernel contrast. Training and validation MSE provide context features. We apply signed $\log(1 + |x|)$ transforms, standardize on calibration data, and fit logistic regression with $C = 1$ and ridge regression with penalty 1. The fixed classification threshold is 0.5. The benchmark rules learn validation benefits under fixed update budgets; separate deployment rules learn validation outcomes from calibration actions with 30-second budgets. The bundled rules record the exact transforms, coefficients, and feature order.

Each benchmark branches into nonlinear, initial-tangent, and refreshed-tangent training. Tangent branches retain their anchor predictor and optimize its linearized parameter increment. The branches share SGD sample streams and update counts. We evaluate validation MSE after 100, 200, 400, and 800 additional updates and choose each branch's checkpoint by validation MSE. Test labels only score the selected predictions. Full-batch and batch-size-32 configurations share this protocol. At minibatch-trained probes, the velocity diagnostics use the full-batch training gradient field. The deterministic flow theorem therefore supplies a local reference quantity, not a theorem for these stochastic trajectories.

I.2 ONLINE DECISION BUDGETS

The deployment study uses eight configurations: two architectures for each of four real-data families, all with minibatch size 32. Six policies share each case: continue, freeze, validation race, combined diagnostics, signed contrast with MSE context, and kernel drift with MSE context. Continue and freeze both select their final checkpoint by ordinary validation MSE. All policies may retain the probe checkpoint.

Each policy receives 30 seconds, including prefix training, diagnostics, tangent construction, subsequent training, and validation selection. We exclude data/model setup and final test scoring equally. Each policy makes five validation assessments on the same examples: one at the probe and four later. The race divides the remaining budget equally between continuation and freezing, uses two later assessments per arm, and alternates arm order by seed. Other policies space their four assessments across the remaining budget. Diagnostic derivatives use more information than scalar validation MSE, despite the matched examples and assessment counts; the timer charges their computation. The comparison excludes the offline cost of calibration. The largest recorded budget overrun is 1.78 ms across all held-out policy jobs.

Policy	MSE reduction	95% interval
Always freeze	0.00264	$[-0.02743, 0.03561]$
Combined selector	-0.00032	$[-0.00267, 0.00176]$
Signed selector	-0.00060	$[-0.00359, 0.00113]$
Drift selector	0.00026	$[-0.00198, 0.00327]$

Table 7: Online policy utility over 120 paired cases from four real-data families. Positive values mean lower test MSE than continued nonlinear training with validation checkpoint selection. Intervals resample datasets and shared seed blocks; none establishes a positive aggregate gain.

All three learned selectors choose continuation in all 120 cases. Calibration favored continuation in 81 of 96 paired cases. The combined selector's held-out continuation probabilities range from 0.559 to 0.989; none crosses the fixed threshold. Timing overhead and different available checkpoints can still change its selected MSE relative to the ordinary continuation baseline. Always freezing improves mean MSE by 0.04877 on Fashion-MNIST but worsens it by 0.03822 on wine. Every policy retains the probe on every CIFAR and diabetes case. A denser early checkpoint schedule could change the comparison; these results do not rule out a useful controller under another protocol. We evaluated the raw signed contrast as a predictor only and did not deploy it or a context-only policy.

I.3 Statistical units and sensitivity

The primary held-out cohort contains 256 benchmarks and 384 policy runs over eight seeds. The complete reserve adds 224 benchmarks and 336 policy runs over seven fresh seeds, giving 15 seeds per configuration. We retain the final incomplete seed's 32 benchmarks and 45 policy runs in the raw record but exclude them from balanced summaries, following the preset completeness rule. Thus 720 policy runs represent 120 paired cases, not 720 independent decisions.

Seeds and configurations share datasets. For uncertainty, we resample dataset families and then shared seed blocks, retaining configurations that share a seed together. Policy estimates weight the four dataset families equally. Benchmark accuracy and AUC in Table 2 pool the 480 cases; they use different weights from the task-balanced comparisons. The raw signed rule's task-balanced accuracy gain over MSE context is 0.224, with 95% interval $[-0.028, 0.483]$ across six families. This wide interval does not establish broad transfer. These diagnostic comparisons are exploratory and do not carry a multiplicity-adjusted confirmatory claim.

The raw signed contrast's primary and reserve sign accuracies are 60.94% and 54.46%. The combined policy's MSE reduction relative to continuation is -0.000883 $[-0.003089, 0.000000]$ in the primary cohort and 0.000322 $[-0.003358, 0.005384]$ in the reserve. Neither cohort establishes a reproducible decision advantage.

For matched-change examples, we compare pairs within dataset, architecture, and batch regime using symmetric relative magnitude difference at most 10%. Pairs overlap and therefore do not constitute independent trials. Kernel-drift pairs show opposite signs in 186/666 cases; representation-change pairs do so in 254/761. Requiring both absolute benefits to exceed 0.001 leaves 140/532 and 186/609, respectively. In diabetes MLP3 minibatch runs, seeds 100 and 101 have representation changes 0.168797 and 0.168744, but benefits $+0.003502$ and -0.007205. CIFAR pair 0/1 MLP2 full-batch seeds 104 and 114 have changes 0.048028 and 0.047987, with benefits $+0.011335$ and -0.023866. These examples concern predictive comparisons and do not identify a causal effect of representation change.

I.4 Execution and numerical amendment

The run lasted 32,410.9 seconds, from September 20, 2026 at 18:00:54 UTC to September 21 at 03:01:05 UTC. It completed 1,565 valid jobs, including calibration and the partial reserve, with no unresolved failures. Seventeen earlier attempts failed when CIFAR MLP tangent branches diverged. A training-only stability check found largest kernel eigenvalue 142.8897, giving learning-rate product $0.03\lambda = 4.2867$, above the linear squared-loss stability limit 2 under the implementation's normalization. We reduced the learning rate from 0.03 to 0.003 for every CIFAR MLP2 configura-

tion and arm, including both class pairs, both batch regimes, and all seeds. The product then became 0.4287. Other configurations kept learning rate 0.03.

This amendment followed rule freezing. We based it on training stability, retained the failed attempts and original design, and did not tune on held-out test performance or change the fitted rules. The active design identifies the amended configurations explicitly. No further failures followed the repair. A separate analysis correction changed the bootstrap to preserve shared seed blocks; it did not alter measurements, decisions, or frozen rules. The source bundle includes the protocol, amendments, audit, original and active designs, and frozen rules alongside the aggregate results.

J COMPUTATIONAL CONTROLS AND OPTIMIZER COMPARISONS

Two different computational questions. A method can reproduce a prescribed gradient-descent trajectory more cheaply without reaching useful predictions sooner than another optimizer. We tested both questions. For trajectory reproduction, an embedded correction transports the nonlinear residual-gradient defect through cached linear dynamics. It rejects blocks when an embedded prediction-error estimate exceeds its tolerance and restarts from the original anchor when the scaled kernel eigenvalue reaches one. This empirical guard does not certify nonlinear stability. The method combines standard numerical-integration ideas; we do not claim that the local velocity theorem proves its finite-horizon accuracy.

On 216 fixed dataset/architecture/seed combinations, the correction reproduces all saved checkpoints of a 1,600-update, full-batch GD continuation within normalized prediction RMS error 0.005. It saves 53.5% of aggregate computation time (97.5% seed-block bootstrap interval: 51.1% to 56.2%); the median speedup is $2.60\times$ relative to the fastest tested GD implementation per case. Aggregate times are 16.00 and 34.42 seconds, so the aggregate speed ratio differs from the median paired ratio. The implementation comparison includes analytic gradients, ordinary autograd, in-place updates, and `torch.optim.SGD`. All use full batches without momentum. We charge kernel construction, monitoring, validation, and restart work, but exclude common data/model construction and the common 25-update prefix. The nine fixed families include seven real datasets and two synthetic regression families. These intervals describe variability across seeds for that fixed task set.

The median corrected run uses 63 true gradient evaluations and 22 kernel constructions instead of 1,600 gradient evaluations. It therefore saves repeated derivative work when the requested output is a faithful continuation. The same experiment does not justify continuing training after validation has already selected an earlier checkpoint, or establish faster useful training than Adam. The broader benchmark in Appendix K, which includes larger training sets and widths and charges common setup, does not reproduce an aggregate speed advantage. The earlier saving therefore applies to its narrow continuation protocol.

Expanded optimizer search. We ran 24,564 additional optimizer/configuration trials, including development sweeps and implementation checks, across 13 families: wine, diabetes, digits, ionosphere, sonar, breast cancer, spambase, airfoil, concrete, energy, yacht, and two synthetic regression families. Trial count does not equal the number of independent datasets. Most studies use two or three hidden layers of width 16, with caps of 96 training, 48 unlabeled-control, 48 checkpoint-selection, and 128 test examples. A separate study uses width 8 and 192 to 1,024 training examples on airfoil, concrete, energy, and yacht. We use one CPU thread, float64 arithmetic, full-batch squared loss, and a primary wall-clock budget of 0.2 seconds. Each method selects among checkpoints actually available within that budget.

The Adam search expands to 162 settings: nine learning rates, three first-moment decays, two second-moment decays, and three coupled weight-decay values. A prescribed boundary extension adds learning rates when necessary. We also compare tuned GD/momentum, L-BFGS, damped Gauss–Newton/Levenberg–Marquardt, and uncorrected block updates. Preconditioned and regularized corrections constitute separate algorithmic extensions. Their blockwise changing preconditioners do not satisfy the original theorem's constant-preconditioner assumption globally. Development validation selects all settings; fresh splits and initializations supply confirmation results. Test labels never select settings or checkpoints.

An initial 40.2% yacht MSE reduction for a preconditioned correction disappears after stronger Adam tuning. A later regularized correction shows a 13.4% reduction on 48 yacht cases, but its prospectively adjusted interval includes zero. Its unchanged 192-case replication gives mean normalized test MSE 0.009181, versus 0.009813 for the current Adam setting and 0.009031 for another previously selected Adam setting. The correction therefore does not establish a robust advantage. Among 99 cases where the correction and the current Adam setting both reach a frozen validation target, the median Adam/correction time ratio is 0.243; the correction runs faster in only two. We retain failed and divergent trials and allocate a fixed error budget across sequential research stages. Separate reports specify each stage's algorithm, settings, and confidence level.

A direct frozen-kernel fit. We separately test a fitted tangent predictor, using established kernel and linearized-network ideas (Jacot et al., 2018; Lee et al., 2019). For initial parameters θ_0, we center the feature map $\phi(x) = J_{\theta_0}(x)$ using training rows and fit

$$\min_{a,b} \frac{1}{2N}\|y - p(X) - b - \Phi_c a\|^2 + \frac{\lambda}{2}\|a\|^2.$$

Here a is the tangent-feature coefficient vector, b is an unregularized scalar intercept, Φ_c stacks the centered training feature rows, λ is the ridge penalty, and p is either zero or the initial network prediction. One eigendecomposition yields a path of 16 penalties from 10^2 to 10^{-8}; validation selects the penalty. A development pilot fixes the prior to zero on diabetes and to the initial network prediction on ionosphere. Linear ridge and last-hidden-layer random-feature ridge use the same penalty path and provide controls. Large fitted coefficients need not approximate the nonlinear network or its trajectory; this experiment tests a predictor rather than a local-flow approximation.

We retune Adam on six development splits per task using all 162 settings. Before confirmation, we freeze settings, 192 new split/initialization seeds per task, and a validation target of 1.1 times the median development validation MSE of the selected Adam setting. Each case caps this target at 0.8 times its initial validation MSE. The stopping baseline actually stops when it reaches that target or exhausts 0.2 seconds. We also evaluate full-budget Adam with validation checkpoint selection. Adam receives up to 33 validation assessments and ridge receives 17. We count all algorithm-specific fitting and prediction work and include every case in the summaries, whether it reaches the target or not.

	Diabetes		Ionosphere	
Method	Test MSE	Time (ms)	Test MSE	Time (ms)
Adam, full budget	0.594656	200.060	0.484599	200.065
Adam, target stopping	0.621134	58.654	0.495665	72.659
Frozen NTK ridge	0.576004	1.153	0.485265	2.025
Random-feature ridge	0.604402	0.417	0.638176	0.420
Linear ridge	0.556491	0.295	0.538001	0.364

Table 8: Means over all 192 fresh splits/initializations per task. Test MSE uses training-target second-moment normalization. Timing uses serial single-thread CPU execution and includes validation selection. The full-budget baseline spends its allocated budget; the target-stopping baseline measures an actual stopping policy.

The primary criterion requires the frozen-NTK fit to stay within a 5% test-MSE noninferiority margin against both Adam baselines and save at least 20% of mean time against target-stopping Adam. Each paired bootstrap interval uses 200,000 resamples and a per-task error allocation of 0.0001953125 from the sequential budget. Diabetes passes all criteria. Its mean fitting time is 98.0% lower than target-stopping Adam's, and its mean test MSE is 3.1% lower than full-budget Adam's. The quality test establishes the stated noninferiority margin; the mean reduction is a descriptive estimate. The NTK fit reaches the frozen validation target in 120 of 192 diabetes cases, versus 137 for stopped Adam and 139 for linear ridge. Thus the timing result concerns average test quality and fitting cost, not equal validation-target attainment. Ionosphere fails the quality criterion against full-budget Adam.

A further check uses the earlier lean Adam implementation, which performs finite-parameter checks only at checkpoints. Its mean test MSE is 0.598900 on diabetes and 0.489716 on ionosphere. Dia-

betes retains noninferiority, while ionosphere still fails. All saved metrics agree with direct recomputation from prediction vectors. A directional-derivative implementation avoids materializing a Jacobian during inference, but frozen-NTK inference still takes about $1.38\times$ the ordinary network forward-pass time on diabetes in the separate batch benchmark.

Ordinary linear ridge beats the frozen NTK fit in both MSE and fitting time on diabetes (Table 8). The result therefore supports trying a simple regularized predictor before nonlinear training. It does not establish a unique benefit of kernel velocity, the correction, or the signed decision rule. Fresh seeds reuse finite benchmark datasets: a test row in one split may appear in another split's development training data. The intervals quantify repeated-split variability conditional on these datasets, not transfer to independently collected populations. These small CPU models also leave GPU, minibatch, and large-model advantages unresolved.

K Endpoint mechanisms and trajectory fidelity

Design and statistical units. A fixed 8.5-hour campaign completed 42 balanced seed blocks, with seeds 940000 through 940041. Each block crosses eight regression datasets (diabetes, airfoil, concrete, energy, red wine quality, abalone, auto-MPG, and combined-cycle power plant), tanh and softplus activations, hidden widths 16 and 64, training sizes 96 and 256, and initial scaled steps 0.1 and 0.4. Networks have two hidden layers and scalar outputs. The resulting 5,376 cases include 2,688 primary cases on the last four datasets, which did not enter method calibration. All computations use CPU float64 and one Torch thread. We remove exact duplicate rows in the four new datasets and fit feature and target normalization on training rows only. Auto-MPG excludes six rows with missing horsepower and drops car names; abalone uses fixed one-hot sex indicators; power plant uses one workbook sheet. The archive records source URLs, parsing, checksums, and split hashes.

Each seed supplies disjoint training, unlabeled-control, validation, and test sets, with 32, 32, and 64 held-out examples respectively. If necessary, we cap the training size to preserve these sets. We choose the common step size as $\eta = \min\{0.1, s/\lambda_{\max}(G_0)\}$, where $s \in \{0.1, 0.4\}$ and $G_0 = J_0^\top J_0/N$ is the initial Gauss–Newton matrix. Every method pays for the common spectral setup and starts at initialization. We run 1,600 prescribed GD updates and save checkpoints every 100 updates, including zero. The 61,824 saved method trajectories include all methods and both primary and bridge families; comparisons within a seed block are dependent.

Exact endpoint accounting. At reference updates 0 and 800, we initialize nonlinear and tangent continuations from the same parameters and predictor. We save test prediction vectors after 5, 20, and 100 updates. For an anchored horizon h, let e_h be the nonlinear-minus-tangent prediction difference, r_h the tangent residual, and $u_h = e_h/\|e_h\|$ when $e_h \neq 0$. The signed margin

$$S_h = \frac{2r_h^\top u_h + \|e_h\|}{\sqrt{N_V}}$$

has the same sign as the MSE gap in Equation (17). For early and late horizons $s = 5$ and $\ell = 100$, elementary expansion gives

$$S_\ell - S_s = \underbrace{\frac{(r_s + r_\ell)^\top (u_\ell - u_s)}{\sqrt{N_V}}}_{\text{prediction direction}} + \underbrace{\frac{(r_\ell - r_s)^\top (u_\ell + u_s)}{\sqrt{N_V}}}_{\text{remaining error}} + \underbrace{\frac{\|e_\ell\| - \|e_s\|}{\sqrt{N_V}}}_{\text{magnitude}}. \tag{22}$$

This symmetric endpoint decomposition avoids choosing an order for the two bilinear changes. It describes contributions to a sign margin, not additive contributions to the raw MSE difference. We call an endpoint comparison resolved when both absolute MSE gaps exceed 0.001. Among 10,752 anchored comparisons, 2,922 satisfy this criterion and 65 change from negative to positive. The largest absolute term in Equation (22) is prediction direction in 37 cases and remaining error in 28; magnitude is never the largest in this subset. These 65 reversals span 28 seed identifiers and 45 dataset-seed pairs. Their decomposition residuals are at most 4.45×10^{-16}. Repeated configurations and anchors do not provide independent evidence for these frequencies. Some other tangent continuations become numerically unstable under the prescribed step size; we retain those cases rather than interpret large raw gaps as useful adaptation. The decomposition supplies an explanation of observed changes, not a causal intervention or a prospective sign predictor.

Method	Pass (%)	Worst RMS	Time (s)	Restarts
Reference GD	100.0	0.0000	0.3204	0
Transported correction	100.0	0.0025	0.3208	316
Adaptive RK	93.8	0.9047	0.2414	0
Without gradient defect	100.0	0.0049	1.8572	323
Without transport	63.1	0.0532	0.2869	317

Table 9: Completed primary trajectory benchmark: 2,688 paired cases from four new task families. Pass means maximum normalized RMS prediction discrepancy at most 0.005. Time includes setup, monitoring, rejected work, and restarts. All rows use the same case set; lower time does not establish an advantage unless fidelity also holds.

Fidelity and full computation cost. We also approximate the prescribed GD trajectory using the transported correction in Appendix J. The primary tolerance is 0.005 for the maximum, over saved checkpoints, of the RMS prediction difference from reference GD on all 128 control, validation, and test inputs, divided by training-target RMS. The algorithm uses training and unlabeled-control inputs for embedded error checks; test outcomes enter only the offline fidelity audit. A spectral or numerical guard triggers a full restart from initialization with exact GD. We charge rejected blocks, diagnostics, initialization, and replay. Fidelity therefore assesses the complete guarded procedure, including its fallback, rather than the correction alone.

Calibration uses 16 cases from the four bridge datasets, disjoint seeds 930000 and 930001, width 16, training size 96, and scaled step 0.1. It selects among safety multipliers 4, 16, and 64; fixed-refresh intervals 4, 10, and 20; and coarse-step factors 2, 4, and 8. Settings remain fixed on the new-family panel. The corrected method starts with block length 20, halves rejected blocks, and can grow to 100. These choices preclude a tuning-free claim. The archive includes the frozen calibration and all comparators.

Table 9 separates approximation accuracy from cost. The guarded correction passes every primary case, including 316 exact-GD restarts, but its mean total time is slightly higher than reference GD. Adaptive Runge–Kutta (RK) is faster on average but fails the tolerance in 6.2% of cases. Removing the gradient defect preserves fidelity at substantially greater cost; removing transport loses fidelity. These ablations support the correction's role in this implementation, without establishing a distinct integration principle. Alternative in-place and autograd GD implementations also exceed the reference-based tolerance in some cases, so nominally identical update rules do not guarantee numerically interchangeable long trajectories under this protocol. We report the reference implementation explicitly rather than assume exact agreement across implementations.

The prospectively specified practical criterion required at least 95% fidelity passes, no discrepancy above 0.05, and at least 20% lower end-to-end time than both tested GD and the development-selected qualifying numerical comparator. The study does not meet this criterion. A paired seed-block bootstrap against the per-case fastest tested GD gives an aggregate time saving of -1.26%, with interval $[-3.41\%, 0.84\%]$ at the allocated error probability 0.0001953125 (200,000 resamples). This interval conditions on four fixed task families and only 42 seed blocks. We make no population-wide efficiency claim, no claim of improved accuracy over Adam, and no claim that the embedded error estimate supplies a global error certificate.

L Discrete dynamics, diagnostic specificity, and non-centered controls

Completed design. A separate fixed-duration study completed 55,200 configurations in 345 balanced seed blocks, with 8.007 hours of case execution. Each configuration evaluates three step scales, giving 165,600 full-batch trajectory pairs. We vary five tasks (a synthetic tanh teacher, diabetes, airfoil, red wine quality, and abalone), two activations, widths 16 and 64, exponents $\alpha \in \{1/2, 1\}$, centered and ordinary outputs, and scalar and layer-diagonal metrics. Each network has one hidden layer and no biases. The layer metric multiplies hidden-weight updates by 0.25 relative to readouts, in addition to the common factor $m^{2\alpha-1}$. Readouts have mean 0.5 and unit

	Abalone			Wine quality		
Diagnostic	8	128	1,024	8	128	1,024
Controls only	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Parameter speed	-0.2221	+0.0646	+0.0069	-0.1140	+0.0168	+0.0063
Feature speed	-0.2077	+0.1398	+0.0483	-0.1007	+0.0018	+0.0016
Kernel velocity	-0.3972	+0.0480	+0.0206	-0.1974	-0.0062	+0.0046
Residual action	-0.4243	+0.0783	+0.0165	-0.2090	+0.0050	+0.0054
Prediction speed	-0.2065	+0.0190	-0.0493	-0.0895	+0.0044	-0.0289
Discrete coefficient	-0.4635	+0.0459	+0.0049	-0.2657	-0.0038	+0.0016

Table 10: Change in held-out squared error for predicting the *log prediction-gap RMS*, relative to controls alone. Each rule adds the named diagnostic to the same controls. Columns give update horizons; negative values favor the diagnostic. Each task contributes 11,040 fresh configurations in 345 paired seed blocks at scaled step 0.1. These values measure forecast error, not predictive test MSE.

variance; hidden weights have variance $1/d$. The softplus variant is $\log(1 + e^z) - \log 2 + 0.1$, with the numerical threshold described in Appendix G; tanh is unchanged.

We use 64 training, 32 validation, and 64 test examples, train-only normalization, and fresh main seeds 970000 through 970344. We pair particles and splits across conditions and use nested width prefixes. The three full-batch step sizes are $\eta = s/\lambda_{\max}(K_0/N)$ for $s \in \{0.02, 0.1, 0.4\}$. Checkpoints are $0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 1600$. We retain 200 divergent trajectories, all at scaled step 0.4, and exclude those trajectories from completed-trajectory summaries. The other 165,400 trajectories complete. Failures occur on abalone (46), airfoil (15), diabetes (42), wine quality (97); the fixed step-0.1 diagnostic comparison has no failures. The archive contains the complete failure list. Seed-block intervals describe repeated splits and initializations on these fixed datasets; they do not represent 55,200 independent tasks.

Short-horizon approximation and matching. At scaled step 0.1, the median first-update relative prediction-vector error of Equation (16) ranges from 0.0193 to 0.0321 across the five tasks; applying the continuous quadratic term at $t = \eta$ instead gives 0.4561 to 0.9434. At eight updates, the discrete error already ranges from 0.8220 to 1.3826. Thus even a correct local expansion can have poor quantitative fidelity at an ordinary finite horizon. Independent step-halving checks show cubic absolute remainder decay at fixed update counts. Initial kernels match across exponents exactly in the stored arithmetic. Residual-matched velocity scaling has maximum relative discrepancy 6.53×10^{-16}, whereas ordinary non-centered comparisons can violate the simple scaling law because their residuals differ. The scalar-clock control uses $\|\dot{K}_0 r_0\|/[\lambda_{\max}(K_0/N)\|K_0 r_0\|]$; rescaling A by $c > 0$ and η by $1/c$ preserves this rate and the paired predictions in these checks. These controls do not cover Adam's evolving preconditioner.

Incremental diagnostic information. We fit separate ridge log-linear predictors of the RMS nonlinear-minus-tangent prediction gap at horizons 8, 128, and 1,024. Four development seeds (960000 through 960003) on the synthetic, diabetes, and airfoil tasks supply 384 configurations at scaled step 0.1. Every rule includes an intercept, log width, scaling exponent, centering and activation indicators, metric indicator, log step size, and log initial residual RMS. We standardize these controls using development statistics and add one log diagnostic, clipping diagnostic values below 10^{-15}. The response is the log evaluation-gap RMS, clipped below 10^{-12}. Ridge penalty 1 applies to standardized non-intercept coefficients. The rules remain fixed for the new seeds and the abalone and wine tasks. Those two tasks were absent from rule fitting, but appeared in the preceding study; they are not wholly untouched research datasets.

For the training set, the diagnostics are residual action $\|\dot{K}_0 r_0\|/N^{3/2}$, scaled kernel velocity $\|\dot{K}_0\|_F \|r_0\|/N^{3/2}$, parameter speed $\|u_0\|$, hidden-activation speed divided by $\sqrt{Nm}$, and prediction speed $\|K_0 r_0\|/N^{3/2}$. The discrete diagnostic is the RMS of the coefficient of η^2 in Equation (16) at $k = 8$. Its evaluation uses validation and test inputs, but no evaluation labels. The target gap also uses those 96 inputs. All other diagnostics use training data. This is a comparison of departure forecasts, not a test-error or deployment-policy comparison.

Task	Scaled step	Nonlinear	Frozen	Difference	95% interval
Abalone	0.02	0.6376	0.6463	-0.0087	$[-0.0115, -0.0059]$
Abalone	0.10	0.6208	0.6210	-0.0001	$[-0.0030, +0.0027]$
Abalone	0.40	0.6285	0.6244	+0.0041	$[+0.0004, +0.0077]$
Airfoil	0.02	0.5170	0.5310	-0.0140	$[-0.0166, -0.0116]$
Airfoil	0.10	0.5049	0.5134	-0.0085	$[-0.0118, -0.0051]$
Airfoil	0.40	0.5091	0.5171	-0.0080	$[-0.0121, -0.0041]$
Diabetes	0.02	0.6035	0.5984	+0.0051	$[+0.0037, +0.0065]$
Diabetes	0.10	0.6078	0.6030	+0.0048	$[+0.0027, +0.0068]$
Diabetes	0.40	0.6104	0.6055	+0.0048	$[+0.0025, +0.0071]$
Synthetic teacher	0.02	0.0879	0.1150	-0.0271	$[-0.0285, -0.0257]$
Synthetic teacher	0.10	0.0864	0.1160	-0.0297	$[-0.0312, -0.0281]$
Synthetic teacher	0.40	0.0866	0.1167	-0.0301	$[-0.0317, -0.0285]$
Wine quality	0.02	0.8423	0.8378	+0.0045	$[+0.0003, +0.0082]$
Wine quality	0.10	0.8525	0.8518	+0.0007	$[-0.0054, +0.0059]$
Wine quality	0.40	0.8634	0.8594	+0.0040	$[-0.0060, +0.0162]$

Table 11: Validation-selected test MSE for the discrete-dynamics study. Difference is nonlinear minus frozen, so negative values favor nonlinear training. MSE uses training-standardized targets. Intervals resample the 345 seed blocks; within-block configurations remain together. At scaled step 0.4, summaries exclude 200 recorded divergent trajectories.

Table 10 shows a specific short-horizon contribution. At eight updates, residual action reduces log-gap prediction error by 0.4243 on abalone and 0.2090 on wine quality. Descriptive 95% seed-block intervals are $[-0.4578, -0.3917]$ and $[-0.2305, -0.1889]$. Kernel velocity also helps, with reductions 0.3972 and 0.1974. The discrete coefficient gives the largest point reductions, 0.4635 and 0.2657. These estimates do not establish universal superiority over kernel velocity, and the intervals compare each rule with controls rather than compare diagnostics directly. At 1,024 updates, residual action worsens error on both tasks, whereas initial prediction speed helps. The horizon dependence supports the local interpretation of the diagnostic. High raw correlations pooled across scaling settings alone would not establish this incremental information.

Validation-selected outcomes. Each nonlinear and tangent branch separately selects its checkpoint by validation MSE. Table 11 reports absolute test errors and paired differences for all tasks and step scales. Nonlinear training improves the synthetic and airfoil means, worsens diabetes means, and gives mixed abalone and wine results. We report all scales rather than select a favorable one. These are exploratory comparisons with descriptive 95% intervals from 10,000 whole-seed-block resamples; we make no multiplicity-adjusted generalization claim. The step-0.4 summaries condition on completion and must be read with the failure counts above.

A retrospective discrete bound. Let $E_j = f_{X,j} - f^{\mathrm{K}}_{X,j}$ and $F_0 = I - \eta K_0/N$. Define the observed forcing

$$b_j = f_{X,j+1} - f_{X,j} + \eta K_0(f_{X,j} - y)/N.$$

Subtracting the frozen update gives the exact recurrence $E_{j+1} = F_0 E_j + b_j$. Since $E_0 = 0$, induction and the triangle inequality yield

$$\|E_k\| \le \sum_{j=0}^{k-1} \|F_0\|_{\mathrm{op}}^{k-1-j} \|b_j\|. \tag{23}$$

The same inequality holds for RMS norms. The numerical audit finds maximum bound violation 8.19×10^{-17}. At scaled step 0.1 and update 1,024, the median bound-to-gap ratio ranges from 5.60 to 9.57 across tasks. Equation (23) uses the observed trajectory, so it supplies an accounting bound rather than an initialization-only forecast or a cheap controller.

Ordered minibatch expansion. For fixed batches $i = 0, \ldots, k-1$, let g_i and $\mathcal{H}_i$ denote the batch loss gradient and Hessian at θ_0, G_i its Gauss–Newton matrix, and $u_i = Ag_i$. Comparing nonlinear

and affine updates with the same batches gives

$$e_{V,k} = \eta^2 \left[J_{V,0} A \sum_{0 \le j < i < k} (\mathcal{H}_i - G_i) u_j + \frac{1}{2} D^2 f_V(\theta_0) \left[\sum_i u_i, \sum_i u_i \right] \right] + O(\eta^3).$$

To obtain this formula, expand each batch gradient at θ_0: the parameter increment is $-\eta \sum_i u_i + \eta^2 \sum_{j<i} A \mathcal{H}_i u_j + O(\eta^3)$. The affine model replaces $\mathcal{H}_i$ with G_i; output expansion supplies the final curvature term. This argument fixes the batch sequence and k while $\eta \to 0$ and makes no stochastic convergence claim. On the 27,600 width-16 configurations, we check eight shared batches of size 16, saving horizons 1, 2, 4, and 8 and step factors 1, 1/2, and 1/4. Independent step-halving verification confirms the cubic remainder. This extension explains why a full-batch flow coefficient cannot substitute for ordered batch interactions.