NEURAL FLIP LAWS AND TANGENT STABILIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A positive flip coefficient supports a local period-two branch of projected gradient descent (GD), while tangent motion can suppress oscillations in full GD. We quantify this distinction in wide additive networks. Exact empirical functionals determine the projected flip coefficient and the tangent-drift coefficient; two Softplus initialization laws share the same population tangent kernel but produce opposite flip signs. For one residual, balanced i.i.d. initialization with positive tangent drift, and a near-critical step calibrated by the initial kernel, a single GD trajectory tracks an explicit amplitude–sharpness system with $O_P(m^{-1/2})$ error over $O(\sqrt{m})$ steps, converges to its own interpolant, and becomes subcritical under either flip sign. The theorem covers unbounded initializations through a finite-eighth-moment derivative envelope. Experiments from random initialization verify the trajectory law, while endpoint interventions identify the tangent sharpness direction that produces damping.

1 INTRODUCTION

The quadratic stability threshold $\eta\lambda_{\max} = 2$ identifies when an update of gradient descent (GD) first becomes unstable. Near an interpolating minimum, nonlinear terms can support oscillations in a fixed normal direction. Full GD can also move along the manifold of minima and lower sharpness. When does this second motion suppress the oscillations predicted by the first?

Local bifurcation theories describe the first effect through a flip Lyapunov coefficient c_1. Its positive, supercritical sign supports a small stable period-two branch near an isolated minimum or within the fixed normal space of a Morse–Bott minimum manifold (Gan, 2026; Mulayoff & Stich, 2026). These statements condition on the endpoint geometry, while full GD also follows tangent motion that changes that geometry.

Our main result follows a single GD trajectory from random initialization to its own interpolant. For a smooth width-m additive network with one residual, balanced i.i.d. initialization, positive tangent drift, and a finite-eighth-moment derivative envelope, choose a fixed step from the initial kernel with detuning of order $m^{-1/2}$. Then Theorem 6.1 tracks residual amplitude and sharpness with $O_P(m^{-1/2})$ error over $O(\sqrt{m})$ steps and proves subsequent convergence at fixed width. The final endpoint is subcritical for nonstationary initialization under either sign of its projected flip coefficient. The moment condition includes compact and Gaussian initializations for the standard smooth activations considered here.

Two neural quantities make this behavior computable from initialization. For the empirical block law $\widehat{\rho}_m$ at a regular interpolant, we derive the exact identities

$$mc_1(\theta_*) = \mathfrak{F}(\widehat{\rho}_m), \qquad m\,\|\nabla_{\mathcal{M}}\mathsf{S}(\theta_*)\|^2 = \mathcal{D}(\widehat{\rho}_m).$$

The signed functional $\mathfrak{F}$ measures projected flip geometry; the nonnegative $\mathcal{D}$ is the tangent-drift coefficient. Sample-space kernel solves and block derivatives evaluate both. These identities hold with multiple residuals, and lazy i.i.d. gradient flow selects their population values (Theorem 4.4). In the scalar dynamical regime, the direct GD trajectory reaches an endpoint with the same limits.

The population tangent kernel leaves both nonlinear coefficients undetermined. With an independent trainable linear readout, the kernel uses only the second readout moment, whereas the functionals can use the fourth. Two explicit Softplus readout laws have identical population kernels on every input pair and opposite nonzero flip signs, with positive tangent drift in both cases (Corollary 5.3).

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.

Their GD trajectories from initialization converge and retain those opposite signs at their own endpoints (Corollary F.5). This construction separates the linear threshold, projected bifurcation geometry, and full training dynamics within one model.

The proof connects a near-critical transient to convergence. Individual GD updates reverse an order-one residual and cannot be accumulated as small parameter increments. Pairing successive updates cancels their leading normal motion and exposes the tangent drift. The resulting amplitude–detuning system has a conserved ellipse and carries the trajectory into a negative-detuning region. A residual-relative estimate then gives geometric contraction and summable paired displacement, upgrading finite-horizon tracking to fixed-width convergence.

At a common smaller detuning proportional to $|c_1|$, full and projected GD separate on the $\sqrt{m}$ horizon: the projected amplitude stays asymptotically constant, while full GD follows a hyperbolic-secant decay (Corollary 6.2). Initialization-to-interpolation experiments test the endpoint theorem directly. Multiple-residual tests examine the same damping prediction beyond the scalar theorem, and five-arm deep-network interventions isolate the tangent sharpness direction that controls it.

2 Related work

Empirical and reduced EoS dynamics. The edge-of-stability regime was documented across neural architectures by Cohen et al. (2021). Smooth higher-order models subsequently explain self-stabilization, two-step dynamics, and central flows, with experiments using tanh, SiLU/Swish, Softplus, or GELU depending on the setting (Damian et al., 2023; Chen & Bruna, 2023; Cohen et al., 2025). These works motivate our separation of projected flip geometry from tangent self-stabilization.

Earlier bifurcation-based trajectory models address the complementary global dynamics question in deliberately low-dimensional networks. Song and Yun prove initialization-independent trajectory alignment to a bifurcation diagram for a two-layer linear network and a one-neuron nonlinear model, each with one datum (Song & Yun, 2023). Kalra, He, and Barkeshli analyze the fixed points of a two-layer linear one-example model and derive an EoS phase with a period-doubling route to chaos (Kalra et al., 2025). These works follow large-step trajectories in special models. Our exact functionals allow general smooth additive blocks and multiple residuals; the quantitative scalar theorem follows one GD trajectory from i.i.d. initialization through its transient and its own interpolating endpoint.

Local bifurcation and flow frameworks. For isolated positive-definite minima, Mulayoff & Stich (2026) characterize the local period-two transition of discrete GD using third and fourth derivatives. The Morse–Bott framework of Gan (2026) replaces the inverse Hessian by its normal-space pseudoinverse and separates the projected normal branch from tangent drift. Both analyses are local to a specified minimum. Our results add a random-neural selection law and derive the paired full-GD update used in the transfer theorem.

MacDonald and Vidal derive normal–tangent expansions and large-step convergence near a manifold of flat minima (MacDonald & Vidal, 2026). Their main geometric regime assumes local sharpness minima with a positive normal Riemannian Hessian, invariant top fibers, and a positive cubic coefficient. Our scalar transfer assumes a nonzero tangent sharpness gradient and allows either flip sign, with explicit $m^{-1/2}$ error on a $\sqrt{m}$ horizon. A separate residual-relative tail argument proves convergence of the GD trajectory from initialization.

Edge coupling and Edge Flow model sharpness forcing and slow EoS dynamics (Litman, 2026; Marion, 2026). Building on this feedback mechanism, our theorem supplies a width-dependent error bound, explicit selected neural coefficients, and fixed-width convergence.

Higher-order tangent geometry. Neural tangent hierarchies track higher derivatives of smooth wide networks, including Softplus (Huang & Yau, 2020). Second-order regression models connect such nonlinear tangent evolution to progressive sharpening (Agarwala et al., 2023). Other work separates tangent and normal motion or studies NTK eigenvector evolution at EoS (MacDonald et al., 2025; Jiang et al., 2025). We instead combine these local derivative objects into one scalar coefficient and identify its selected population law.

3 SETUP AND THE LOCAL COEFFICIENT

Let $r\colon \mathbb{R}^P \to \mathbb{R}^n$ be a C^4 residual map and

$$L(\theta) \coloneqq \frac{1}{2n}\,\|r(\theta)\|_2^2\,. \tag{1}$$

Fix an interpolant θ_* with $r(\theta_*) = 0$. Write $J \coloneqq Dr(\theta_*)$ and $H \coloneqq D^2L(\theta_*)$. We assume that J has full row rank and that the largest positive eigenvalue λ of H is simple, with unit eigenvector v. Define residual directional jets

$$\begin{aligned} u &\coloneqq Jv, & q &\coloneqq D^2r(\theta_*)[v,v], & t &\coloneqq D^3r(\theta_*)[v,v,v],\\ B_v w &\coloneqq D^2r(\theta_*)[v,w], & z &\coloneqq \frac{1}{n}\big(J^\top q + 2B_v^\top u\big). \end{aligned}$$

At the flip threshold $\eta\lambda = 2$, we use the normal-form convention

$$c_1 \coloneqq \frac{\eta}{2}D^3L[v,v,H^\dagger z] - \frac{\eta}{6}D^4L[v,v,v,v]. \tag{2}$$

With this convention, positive c_1 is the supercritical sign supporting the small stable period-two branch in local bifurcation theory.

The C^4 neighborhood regularity supports the discrete-GD normal form. At interpolation, the squared-loss reduction below uses only the first three residual jets because the fourth residual derivative is multiplied by $r(\theta_*) = 0$.

At a regular interpolant, the implicit-function theorem makes $\mathcal{M} \coloneqq r^{-1}(0)$ a local manifold with $T_{\theta_*}\mathcal{M} = \ker J = \ker H$ and $N_{\theta_*}\mathcal{M} = \operatorname{range}(J^\top)$. Thus the squared loss is Morse–Bott in the normal directions. A simple largest positive Hessian eigenvalue gives one simple normal multiplier -1 at $\eta\lambda = 2$, while all other normal multipliers lie strictly inside the unit circle.

Under these geometric and spectral hypotheses, the local projection result of Gan (2026) interprets $c_1 > 0$ as the supercritical sign for a stable period-two branch of the discrete-GD dynamics projected onto $N_{\theta_*}\mathcal{M}$; the isolated positive-definite counterpart appears in Mulayoff & Stich (2026). The conclusion is local to a specified minimum and a sufficiently small positive detuning $\delta \coloneqq \eta\lambda - 2$. With the convention in (2), its leading normal half-amplitude p_δ satisfies

$$p_\delta^2 = \frac{\delta}{c_1} + o(\delta) \qquad (c_1 > 0).$$

This local result supplies the projected comparison in Corollary 6.2. The direct trajectory theorem, Theorem 6.1, follows full GD from random initialization to its own subcritical interpolant.

Proposition 3.1 (Residual-jet reduction)**.** *At every interpolant satisfying the preceding rank and gap conditions,*

$$\begin{aligned} H &= \frac{1}{n}J^\top J, & D^4L[v,v,v,v] &= \frac{1}{n}\big(3\,\|q\|_2^2 + 4\,\langle u,t\rangle\big)\\ c_1 &= \frac{1}{\lambda}\left(z^\top H^\dagger z - \frac{\|q\|_2^2}{n} - \frac{4\,\langle u,t\rangle}{3n}\right). \end{aligned} \tag{3}$$

The formula separates a nonlocal relaxation term, an aggregate second-jet penalty, and a third-jet contraction. Its sign is governed by cancellations among these three terms rather than by a neuron-wise contribution.

Proposition 3.2 (Sample-space, matrix-free evaluation)**.** *Let $K \coloneqq JJ^\top$, let (μ, e) be its unit top eigenpair, and orient e arbitrarily. Then*

$$\mu = n\lambda, \qquad v = \frac{J^\top e}{\sqrt{\mu}}, \qquad z^\top H^\dagger z = n \left\| K^{-1} J z \right\|_2^2. \tag{4}$$

Consequently, c_1 can be evaluated entirely in sample space using residual JVPs and VJPs, a top eigensolve for K, directional second and third derivatives, and one linear solve.

Our population results use the additive block model

$$f_m(x_i;\theta) \coloneqq \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \psi_i(\theta_j), \qquad r_i(\theta) \coloneqq f_m(x_i;\theta) - y_i. \tag{5}$$

The principal neural specialization is $\psi_i(a,\beta) = as(\beta^\top \widetilde{x}_i)$, where $\widetilde{x}_i = (x_i^\top, 1)^\top$ includes a trainable bias and s is smooth. The exact functional below applies to arbitrary smooth block maps; the ridge structure supplies the explicit same-kernel construction.

4 Two exact neural functionals

For a probability measure ρ on one block, put

$$g_i(\xi) \coloneqq \nabla \psi_i(\xi), \qquad A_i(\xi) \coloneqq \nabla^2 \psi_i(\xi), \qquad T_i(\xi) \coloneqq \nabla^3 \psi_i(\xi),$$
$$K_\rho(i,k) \coloneqq \int \langle g_i(\xi), g_k(\xi) \rangle \, d\rho(\xi).$$

Whenever K_ρ is invertible with unit simple top eigenpair (μ, e), define $h_\rho(\xi) \coloneqq \sum_i e_i g_i(\xi)$ and

$$Q_i(\rho) \coloneqq \frac{1}{\mu} \int A_i[h_\rho, h_\rho] \, \mathrm{d}\rho, \qquad R_i(\rho) \coloneqq \frac{1}{\mu^{3/2}} \int T_i[h_\rho, h_\rho, h_\rho] \, \mathrm{d}\rho,$$
$$\zeta_\rho(\xi) \coloneqq \frac{1}{n} \left(\sum_i Q_i(\rho) g_i(\xi) + 2 \sum_i e_i A_i(\xi) h_\rho(\xi) \right), \quad S_k(\rho) \coloneqq \int \langle g_k(\xi), \zeta_\rho(\xi) \rangle \, \mathrm{d}\rho(\xi). \tag{6}$$

The neural flip functional is

$$\mathfrak{F}(\rho) \coloneqq \frac{n^2}{\mu} \left\| K_\rho^{-1} S(\rho) \right\|_2^2 - \frac{1}{\mu} \left\| Q(\rho) \right\|_2^2 - \frac{4}{3\sqrt{\mu}} \left\langle e, R(\rho) \right\rangle. \tag{7}$$

Its three terms are precisely the relaxation, second-jet, and third-jet terms in (3), but expressed through one-block averages.

Theorem 4.1 (Exact empirical flip functional)**.** *Let $\widehat{\rho}_m \coloneqq m^{-1} \sum_{j=1}^m \delta_{\theta_j}$. At every regular interpolant of* (5)*,*

$$m c_1(\theta_*) = \mathfrak{F}(\widehat{\rho}_m).$$

This deterministic identity applies to every C^4 additive block model at a regular interpolant satisfying the rank and spectral conditions of Section 3.

The tangent-drift coefficient has a parallel representation. Let $\mathsf{S}(\theta) \coloneqq \lambda_{\max}(D^2 L(\theta))$, and let $\nabla_{\mathcal{M}}\mathsf{S}$ denote its tangent gradient at interpolation. Define

$$\chi_\rho \coloneqq \frac{2}{n}\sum_i e_i A_i h_\rho, \qquad U_k(\rho) \coloneqq \int \langle g_k, \chi_\rho\rangle \, \mathrm{d}\rho, \qquad \mathcal{D}(\rho) \coloneqq \int \|\chi_\rho\|^2 \, \mathrm{d}\rho - U^\top K_\rho^{-1} U.$$

Proposition 4.2 (Exact tangent functional)**.** *At every regular interpolant of* (5)*,*

$$m\,\|\nabla_{\mathcal{M}}\mathsf{S}(\theta_*)\|^2 = \mathcal{D}(\widehat{\rho}_m) = \min_{c\in\mathbb{R}^n} \int \|\chi_{\widehat{\rho}_m} - \textstyle\sum_i c_i g_i\|^2 \, \mathrm{d}\widehat{\rho}_m.$$

For any admissible law, $\mathcal{D} \geq 0$, with equality exactly when χ_ρ belongs to the sample-feature span in $L^2(\rho)$.

Both identities turn endpoint selection into a derivative-moment convergence problem.

Assumption 4.3 (I.i.d. gradient-flow selection)**.** Fix n, d and let $m \to \infty$. The initial blocks are i.i.d. from a subgaussian law ρ_0 satisfying $\int \psi_i(\xi)\,\mathrm{d}\rho_0(\xi) = 0$ for every sample i. On the unit enlargement of the initialized blocks, one common random envelope $B_m \geq 1$ bounds the block map and its parameter derivatives through order four, and

$$B_m^{16}/\sqrt{m} \xrightarrow{P} 0.$$

The population kernel K_{ρ_0} has a deterministic positive eigenvalue floor and a simple top eigenvalue separated by a deterministic gap. Every derivative moment entering (6) is integrable, and the fixed-dimensional target vector is bounded in probability. The initial blocks and targets are independent whenever the targets are random.

The appendix proves concentration and endpoint transport for this finite family of derivative moments. The common envelope in Assumption 4.3 gives one concise sufficient condition.

Theorem 4.4 (Joint selected law)**.** *Under Assumption 4.3, additionally assume integrability of the finite family $\langle A_i g_a, A_j g_b\rangle$ for $i, j, a, b \in [n]$. Then the same endpoint satisfies*

$$\Big(mc_1(\theta_{m,*}), m\,\|\nabla_{\mathcal{M}}\mathsf{S}(\theta_{m,*})\|^2\Big) \xrightarrow{P} (\mathfrak{F}(\rho_0), \mathcal{D}(\rho_0)).$$

Appendix D lists and transports the additional moments. Corollary C.2 verifies the assumptions for Softplus with Gaussian hidden parameters, while the scalar GD theorem below supplies direct endpoint transport along its own trajectory.

5 Equal kernels and opposite flip signs

For blocks $\psi_i(A, \omega) = A\varphi_i(\omega)$ with independent readout and hidden laws, write $m_r \coloneqq \mathbb{E}A^r$. The kernel fixes the linear stability threshold. The following moment identities identify the additional information needed for the two nonlinear coefficients.

Theorem 5.1 (Four-moment universality)**.** *The population functional $\mathfrak{F}(P_A \otimes P_\omega)$ depends on the readout law only through (m_1, m_2, m_3, m_4). If*

$$m_1 = m_3 = 0,$$

then the second-jet vector Q and sample-space Riesz vector S in (6) *vanish exactly. The remaining functional uses only (m_2, m_4); symmetry of the readout law is sufficient but not necessary.*

This finite-moment closure is a consequence of trainable-readout linearity: the tangent kernel has degree two in A, the second-jet and Riesz terms have odd degrees one and three, and the third jet has even degrees two and four.

Third-order balance cancels the odd contractions in the flip functional, but leaves a positive square in the tangent functional.

Theorem 5.2 (Readout-moment closure for tangent motion). *Assume $m_4 < \infty$, the required hidden-derivative contractions are integrable, and the population kernel is positive definite with a simple top eigenpair (μ, e). Put $\alpha := \sum_i e_i \varphi_i$, $\beta := \nabla\alpha$, and $\mathsf{H} := \nabla^2\alpha$. The functional $\mathcal{D}$ depends on the readout law only through its first four moments. Under $m_1 = m_3 = 0$,*

$$\mathcal{D}(\rho_0) = \frac{4}{n^2}\mathbb{E}\left[A^2\|\beta\|^4 + \left\|\alpha\beta + A^2\mathsf{H}\beta\right\|^2\right] \geq \frac{4m_2}{n^2}\mathbb{E}\|\beta\|^4.$$

For $m_2 > 0$, equality $\mathcal{D} = 0$ holds exactly when $\beta = 0$ almost surely.

The two formulas allow a controlled change of nonlinear geometry while preserving the population kernel.

Corollary 5.3 (Same population kernel, opposite flip signs). *For Softplus ridge blocks, initialize hidden weights at zero and bias at $\log 3$. Compare readout laws $P^-\{\pm 1\} = 1/2$ and $P^+\{0\} = 15/16$, $P^+\{\pm 4\} = 1/32$. Both have the population kernel*

$$K_\infty(x, x') = (\log 4)^2 + \frac{9}{16}\langle \widetilde{x}, \widetilde{x}'\rangle.$$

For the single training input $x = 0$ and any fixed label, both satisfy the assumptions of Theorem 6.1, with $D_\infty > 0$, but their selected limits of mc_1 have opposite signs. Specifically, with $\ell := \log 4$, $\lambda_\infty = \ell^2 + 9/16$, and $m_4 \in \{1, 16\}$,

$$\mathfrak{F}(\rho_0) = \frac{27(m_4 - 8\ell)}{512\lambda_\infty^2}, \qquad D_\infty = \frac{81}{64} + \frac{9}{4}\ell^2 + \frac{27}{32}\ell + \frac{81}{1024}m_4.$$

The analytic signs follow from $1 < \log 4 < 2$, and kernel equality holds on every input pair. Changing the fourth readout moment changes both $\mathfrak{F}$ and D_∞; Corollary F.5 gives the opposite signs at actual GD endpoints.

6 Tangent self-stabilization on the width scale

The flip coefficient describes a fixed-normal projection. Full GD also moves along the interpolation manifold, changing the stability threshold itself. We now quantify this motion for one residual. Write $f_m := m^{-1/2}\sum_j \psi(\theta_j)$, $L_m := (f_m - y)^2/2$, and $\epsilon := m^{-1/2}$.

Theorem 6.1 (One GD trajectory and its selected coefficients). *Let ψ be C^4 and initialize the blocks i.i.d. from ρ_0. Suppose a measurable envelope $u(\xi) \geq 1$ bounds the first four block derivatives on the unit ball centered at ξ, with $\mathbb{E}_{\rho_0}u^8 < \infty$ and $\mathbb{E}_{\rho_0}\psi^2 < \infty$. For $h = \nabla\psi$ and $A = D^2\psi$, assume*

$$\mathbb{E}\psi = 0, \qquad \lambda_\infty = \mathbb{E}\|h\|^2 > 0, \qquad \mathbb{E}[h^\top A h] = 0, \qquad D_\infty = 4\mathbb{E}\|Ah\|^2 > 0.$$

Fix a finite label y and $d_0 \in \mathbb{R}$. At initialization define $\lambda_0 := \left\|\nabla f_m(\theta^0)\right\|^2$ and $a_{0,m} := (f_m(\theta^0) - y)/\sqrt{\lambda_0}$. On $\lambda_0 > 0$, run full GD from θ^0 with the fixed step $\eta_m := (2 + \epsilon d_0)/\lambda_0$. The comparison functions (a_m, d_m) satisfy

$$a_m' = 2d_m a_m, \qquad d_m' = -4D_\infty a_m^2/\lambda_\infty^2, \qquad (a_m(0), d_m(0)) = (a_{0,m}, d_0).$$

For every fixed T, uniformly for $k \leq T/\epsilon$,

$$\frac{f_m(\theta^{(2k)}) - y}{\sqrt{\lambda_0}} = a_m(\epsilon k) + O_P(\epsilon), \qquad \frac{\eta_m \mathsf{S}(\theta^{(2k)}) - 2}{\epsilon} = d_m(\epsilon k) + O_P(\epsilon).$$

With probability tending to one, the same iterates converge to their own regular interpolant $\theta_{m,\infty}$, block i moves $O_P(\epsilon u(\theta_i^0)^2)$ over the entire trajectory, and at that endpoint

$$mc_1 = \mathfrak{F}(\rho_0) + O_P(\epsilon), \qquad m\left\|\nabla_{\mathcal{M}}\mathsf{S}\right\|^2 = D_\infty + O_P(\epsilon).$$

If $\mathrm{Var}(\psi) + y^2 > 0$, its normalized final sharpness is

$$\frac{\eta_m \mathsf{S}(\theta_{m,\infty}) - 2}{\epsilon} = -\sqrt{d_0^2 + 2D_\infty a_{0,m}^2/\lambda_\infty^2} + O_P(\epsilon).$$

Otherwise initialization already interpolates and GD is stationary.

Proof sketch. Put $r \coloneqq f_m - y$, $g \coloneqq \nabla f_m$, $H \coloneqq D^2 f_m$, $\lambda \coloneqq \|g\|^2$, and $d \coloneqq (\eta_m \lambda - 2)/\epsilon$. Fix a failure probability $\alpha > 0$. Concentration and tightness localize the initial residual, derivative averages, kernel floor and drift to deterministic bounds depending on α. Balance gives $g^\top H g = O_\alpha(\epsilon^2)$ throughout a stopped neighborhood. Expanding two successive updates cancels the leading normal increments and gives

$$\theta^{++} - \theta = \eta_m \epsilon d r g - \eta_m^2 r^2 H g + O_\alpha(\epsilon^2 r^2).$$

The first term changes the residual; the second moves predominantly tangentially. The weighted expansion in Lemma F.2 yields

$$r^{++} = (1 + 2\epsilon d) r + E_a, \qquad |E_a| \le C_\alpha \epsilon^2 |r|,$$
$$d^{++} - d = -\frac{4 D_\infty}{\lambda_\infty^3} \epsilon r^2 + E_d, \qquad |E_d| \le C_\alpha \epsilon^2 (|r| + r^2).$$

These are the Euler updates for the stated system after normalizing r by $\sqrt{\lambda_0}$. Discrete Gronwall gives fixed-time tracking. The rank-one decomposition $D^2 L = g g^\top + r H$ and balance give $\mathsf{S} - \lambda = O_\alpha(\epsilon^2)$, so the same normalized law holds for actual sharpness.

The ODE invariant is $d^2 + 2 D_\infty a^2 / \lambda_\infty^2$. On a localized nonstationary event, it carries the trajectory to $d < -3\beta$ and $|r| < \rho/2$ for fixed $\beta, \rho > 0$. Tracking supplies entrance of GD to this region. The residual-relative bound above then implies

$$|r_{k+j}| \le (1 - \beta\epsilon)^j |r_k|, \qquad \sum_{j \ge 0} |r_{k+j}| \le \frac{\rho}{\beta \epsilon}.$$

The summable detuning error preserves the negative gap; paired block increments sum to $O_\alpha(\epsilon\rho/\beta)$ (Lemma F.3). Thus the iterates converge at fixed width. Summing the invariant error gives the final sharpness formula, and the weighted block-displacement estimates transport the coefficient moments to this same endpoint with $O_P(\epsilon)$ error. The full proof in Appendix F treats arbitrarily small initial residuals by localization and exact interpolation separately. □

The step uses only the initial kernel. Lemma F.4 shows that gradient-flow-endpoint calibration differs by $O_P(m^{-1})$.

Damian et al. (2023, Section 4.2) derive amplitude–sharpness feedback with a conserved potential. Their trajectory theorem assumes positive progressive sharpening; our ellipse arises in the zero-forcing specialization. Our proof derives paired updates for balanced wide blocks and establishes a separate residual-relative tail. The feedback coefficient is computed from initialization, and the result covers both signs of c_1.

Corollary 6.2 (Projected and unprojected GD at a common step)**.** *Under Theorem F.1, fix $a_0 \neq 0$ and instead choose $\eta_m \lambda_* - 2 = |c_1(\theta_*)|\,(2a_0)^2$. Run both full GD and GD projected onto the affine line $\theta_* + \mathbb{R} v_*$, with the same initial perturbation and step. If p_{2k} is the projected amplitude, then for every fixed T, uniformly for $k \le T/\epsilon$,*

$$p_{2k} = a_0 + O_P(\epsilon), \qquad \frac{f_m(\theta^{(2k)}) - y}{\sqrt{\lambda_*}} = a_0 \operatorname{sech}\left(\frac{2\sqrt{2 D_\infty}}{\lambda_\infty} |a_0| \epsilon k\right) + O_P(\epsilon).$$

This distinction holds for either sign of the flip coefficient.

Thus the projected cubic term, which is $O(m^{-1})$, acts more slowly than the leading tangent feedback in this regime. Tangent feedback can therefore suppress the projected period-two behavior in full GD.

7 Tests of tangent stabilization

Endpoint interventions compare projected and full GD with identical settings; projection retains the entire fixed normal space, including weaker modes. Separate runs from random initialization test the single-trajectory endpoint theorem.

Equal kernels permit opposite local dynamics. The scalar test uses both Softplus laws of Corollary 5.3, widths $256, 1024, 4096$, three seeds, target 0.5, $a_0 = 0.2$, and pair-time horizon $T = 3$. All

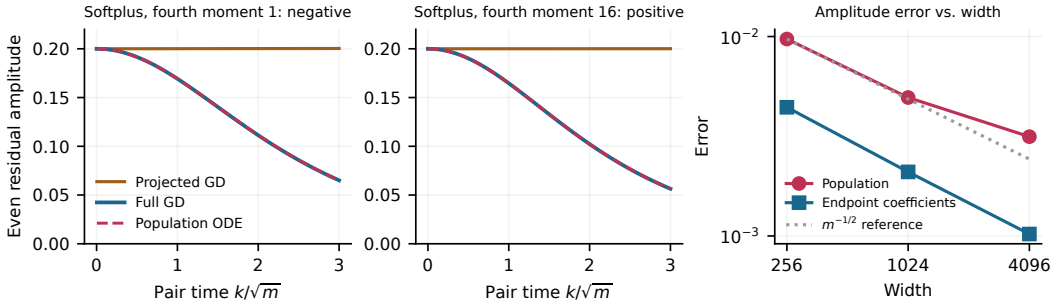

Figure 1: Same population kernel, opposite flip signs, and suppression under both signs. Left/middle: the two exact Softplus laws of Corollary 5.3, width 4096 and seed 20260908, with common projected/full-GD detuning $|c_1|(0.4)^2$ within each panel. Right: largest sampled amplitude errors over four activations and three seeds under $0.3/\sqrt{m}$ detuning. The dotted line is a reference rate; the ODE coefficients are computed independently of the trajectories.

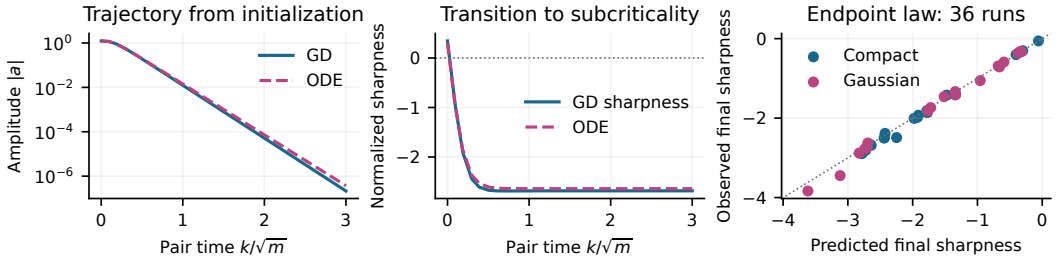

Figure 2: Single-trajectory test of Theorem 6.1. Left/center: one compact width-4096 trajectory from random initialization (seed 20260933, $d_0 = 0.3$); solid curves are GD and dashed curves are the population ODE. Right: observed versus predicted final normalized sharpness for all 36 compact/Gaussian runs; the dotted line is equality. Every run reaches residual 10^{-9}.

18 i.i.d. gradient-flow endpoints reach residual tolerance 10^{-8} and match the analytic flip sign. Under $\delta = |c_1|(0.4)^2$, the largest sampled population-ODE amplitude error decreases from 0.000812 to 0.000197. Full GD damps under both signs while projected amplitudes remain near 0.2 (Figure 1), as in Corollary 6.2.

One trajectory from initialization. All 36 compact/Gaussian runs and all 18 runs of the exact two-sign witness reach residual 10^{-9} using a single fixed-step GD trajectory whose step is chosen from its initialization. All 18 witness endpoints match their predicted flip signs. The GF calibration reference is evaluated post hoc; Appendix G.3 gives settings, errors, and all outcomes.

Width scaling with multiple residuals. We use Gaussian inputs in $d = 16$, independent tanh-teacher targets, Softplus/tanh, five seeds per cell, and $n = 16, 64, 256$ at widths $512, 2048$, plus $n = 64, m = 8192$. L-BFGS with Newton residual corrections reaches all 70 endpoints at RMS tolerance 10^{-6}. Both detunings are tested with $a_0 = 0.2$ and $T = 3$. The endpoint-coefficient ODE is an empirical prediction here: amplitude is $e_*^\top r/\sqrt{\mu_*}$ for the top kernel eigenpair, and errors are maxima over at most 49 sampled even iterates. Its coefficients are computed independently from endpoint derivatives. Table 1 shows the width improvement and the larger errors at $n = 256$.

Overall, 130/140 comparisons complete, with smaller final full-GD amplitude in every completed pair. All ten escapes occur for Softplus at $n = 256$ under coefficient-based detuning: $\delta \in [1.29, 11.93]$ is outside a small-detuning regime. At $n = 256, m = 512$, only 3/5 Softplus runs under $0.3/\sqrt{m}$ and 1/5 tanh runs under coefficient-based detuning meet amplitude-error tolerance 0.02. This experiment extends the scalar prediction empirically to multiple residuals.

A direction-specific intervention in deep networks. We use 256 examples (128 per class) from the scikit-learn copy of the UCI Optical Recognition of Handwritten Digits dataset (Alpaydin & Kaynak, 1998), restricted to digits 0 and 1, with three 128-wide hidden layers and 41,473 pa-

Table 1: Multiple-residual amplitude prediction under $\delta = 0.3/\sqrt{m}$: median [minimum, maximum] sampled error over five seeds, in units of 10^{-3}. All displayed cases complete; initial amplitude is 0.2. The dash denotes a configuration outside this grid.

n	Activation	$m = 512$	$m = 2048$	$m = 8192$
64	Softplus	3.70 [3.06,6.27]	1.32 [1.13,1.75]	0.557 [0.533,0.618]
64	tanh	2.94 [2.79,3.69]	1.24 [1.23,1.32]	0.587 [0.582,0.590]
256	Softplus	19.4 [15.6,23.1]	11.2 [10.1,15.3]	—
256	tanh	8.44 [7.18,14.0]	2.31 [2.02,3.48]	—

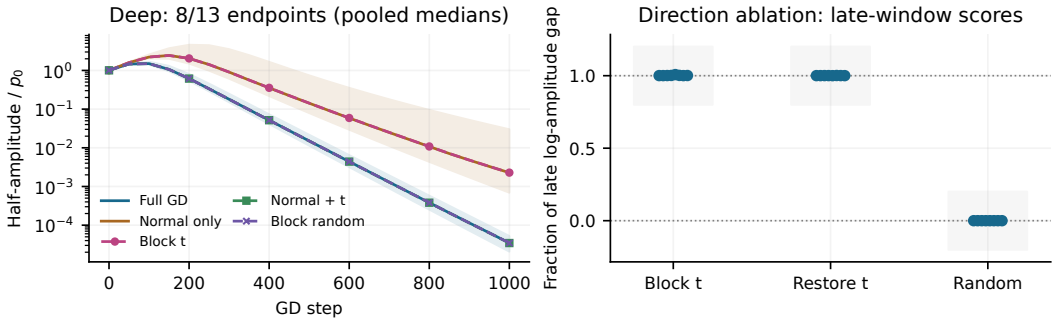

Figure 3: A direction-specific deep-network intervention. Left: median normalized amplitudes pooled over all eight reached endpoints; bands show the full range for full and normal-only GD. Right: each endpoint's fraction of the late log-amplitude gap explained by blocking or restoring t, with the random-direction control. Shaded intervals mark distance 0.2 from the targets one, one, and zero. Eight of thirteen endpoint attempts reach the interpolation tolerance.

rameters. Adam/L-BFGS with Newton corrections reaches residual 10^{-7} in 8 of 13 attempts across GELU, tanh, and SiLU. Each reached endpoint uses the same perturbation, $\delta = 0.01$, $p_0 = \min\{0.02, \delta/(\eta \|P_T \nabla S\|)\}$, and 1,000 updates per intervention.

Freeze $t = P_T \nabla S / \|P_T \nabla S\|$ and a random unit tangent $\xi \perp t$ at the endpoint. Five updates use

$$\theta^+ = \theta_* + A(\theta - \eta \nabla L(\theta) - \theta_*), \qquad A \in \{I, P_N, I - tt^\top, P_N + tt^\top, I - \xi\xi^\top\}.$$

These retain all directions, only normal directions, all but t, normal directions plus t, or all but ξ, respectively. The observable is the half-amplitude $\left|v_*^\top(\theta_k - \theta_{k+1})\right| / (2p_0)$.

Let ℓ_X be its mean logarithm at six snapshots from steps 750 to 1,000, and $G = \ell_N - \ell_{\text{full}}$. Blocking t explains $(\ell_{\text{block}} - \ell_{\text{full}})/G = 1.0005$–$1.0077$ of this gap; restoring only t explains $(\ell_N - \ell_{\text{restore}})/G = 1.0002$–$1.0014$. The random-control fraction has magnitude below 8.2×10^{-9}. All eight cases lie within 0.2 of one, one, and zero, respectively; scores are reported directly. The largest sampled discrepancy between the corresponding amplitude curves is 0.0323 in initial-amplitude units (Figure 3). At step 1,000, normal-plus-t/full amplitude ratios are 0.9946–0.9987, and block-t/normal ratios are 1.0021–1.0493; normal/full ratios span 33.3–573.4. A single fixed tangent direction therefore accounts for most of the additional damping in these selected networks.

These finite-horizon interventions study optimizer-selected interpolants. Normal-only GD also lowers ambient sharpness at five endpoints, so weak normal modes remain material; the tangent direction explains the additional damping of full GD. Together the scalar, multiple-residual, and deep-network tests probe the same mechanism at increasing model complexity.

Independent dense autodifferentiation/projector checks agree within 4.8×10^{-14}, and trajectory subspace leakages are below 1.2×10^{-15}. Full settings, convergence outcomes, and rank sensitivity appear in Appendices G.2, G.4 and G.4.4.

REPRODUCIBILITY STATEMENT

The appendix gives the proofs, experimental settings, preprocessing, numerical tolerances, and convergence outcomes. Supplementary software and data provide the implementations, execution instructions, fixed seeds, raw trajectories, and numerical validation records. Coefficient evaluations are compared with separate sample-space calculations and, at small scale, direct loss derivatives through order four.

AI USE STATEMENT

Generative AI tools were used to identify and refine the research question, survey and organize related work, develop theoretical models and mathematical claims, assist with proof derivations, design and implement experiments, interpret results, and draft and edit the manuscript, code, and references. AI-assisted content was checked against primary literature, independent algebraic implementations, deterministic numerical reruns, and high-precision or interval calculations where stated. The authors take responsibility for the final manuscript, code, and data produced with the aid of generative AI.

REFERENCES

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v202/agarwala23b.html.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50P49. Dataset. DOI: 10.24432/C50P49.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sIE2rI3ZPs.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

Eric Gan. A bifurcation theory framework for gradient descent on the edge of stability. *arXiv preprint arXiv:2606.15551*, 2026. URL https://arxiv.org/abs/2606.15551.

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*, 2020. URL https://proceedings.mlr.press/v119/huang20l.html.

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2e091b1c71ac0e4b4a4cc39ed3d64fcc-Abstract-Conference.html.

Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VZN0irKnl0.

Elon Litman. The origin of edge of stability. *arXiv preprint arXiv:2604.20446*, 2026. URL https://arxiv.org/abs/2604.20446.

Lachlan MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and Rene Vidal. Convergence rates for gradient descent on the edge of stability for over-parametrised least squares. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d9af4d6ac714626b652da5616ca71f99-Abstract-Conference.html.

Lachlan Ewen MacDonald and René Vidal. Dynamics of gradient descent with large step size near a manifold of flat minima. *arXiv preprint arXiv:2607.08380*, 2026. URL https://arxiv.org/abs/2607.08380.

Pierre Marion. Edge flow: A tractable and predictive continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2606.18080*, 2026. URL https://arxiv.org/abs/2606.18080.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in GD and SGD: Beyond quadratic potentials. In *Proceedings of the Conference on Learning Theory*, 2026. URL https://arxiv.org/abs/2602.14789.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PnJaA0A8Lr.

A Proofs of the exact reductions

A.1 Residual jets

Proof. Differentiate (1). Every term carrying an undifferentiated residual vanishes at $r(\theta_*) = 0$, giving

$$H = \frac{1}{n} J^\top J, \qquad \left(D^3 L[v, v, \cdot]\right)^\sharp = \frac{1}{n}(J^\top q + 2B_v^\top u) = z,$$

and

$$D^4 L[v, v, v, v] = \frac{1}{n}\left(3\,\|q\|_2^2 + 4\,\langle u, t\rangle\right).$$

Since $D^3 L[v, v, H^\dagger z] = z^\top H^\dagger z$ and $\eta = 2/\lambda$, substitution into (2) proves (3). □

A.2 Sample-space evaluation

Proof. The nonzero eigenvalues of $H = J^\top J/n$ are those of $K = JJ^\top$ divided by n. Thus $Ke = \mu e$ implies $v = J^\top e/\sqrt{\mu}$ and $\mu = n\lambda$. Full row rank gives

$$H^\dagger = nJ^\top (JJ^\top)^{-2} J,$$

and therefore $z^\top H^\dagger z = n\left\|K^{-1} J z\right\|_2^2$. □

A.3 BLOCKWISE SCALING

Proof. At $\rho = \widehat{\rho}_m$, the top parameter direction has blocks

$$v_j = \frac{1}{\sqrt{m}} \frac{h_\rho(\theta_j)}{\sqrt{\mu}}.$$

Direct differentiation of (5) gives

$$q = m^{-1/2} Q(\rho), \qquad t = m^{-1} R(\rho), \qquad z_j = m^{-1} \zeta_\rho(\theta_j), \qquad Jz = m^{-1/2} S(\rho). \tag{8}$$

Moreover, $\lambda = \mu/n$. Substitute (8) into (3) and use (4); the three terms become exactly m^{-1} times the three terms in (7). □

B PROOF OF THE READOUT-MOMENT LAW

B.1 FOUR-MOMENT EXPANSION

Proof. Let (μ, e) be the top eigenpair of K_{ρ_0} and define

$$\alpha(\omega) \coloneqq \sum_i e_i \varphi_i(\omega), \qquad \beta(\omega) \coloneqq \sum_i e_i \nabla \varphi_i(\omega).$$

The top block direction is $h = (\alpha, A\beta)$ and

$$K_{\rho_0}(i,k) = \mathbb{E}_\omega \left[\varphi_i \varphi_k + m_2 \left\langle \nabla \varphi_i, \nabla \varphi_k \right\rangle \right]. \tag{9}$$

Moreover,

$$\begin{aligned} D^2 \psi_i[h,h] &= 2A\alpha \left\langle \nabla \varphi_i, \beta \right\rangle + A^3 D^2 \varphi_i[\beta, \beta], \\ D^3 \psi_i[h,h,h] &= 3A^2 \alpha D^2 \varphi_i[\beta,\beta] + A^4 D^3 \varphi_i[\beta,\beta,\beta]. \end{aligned} \tag{10}$$

Thus Q uses only m_1, m_3, while R uses only m_2, m_4. To check the remaining Riesz term, put

$$\begin{aligned} U^{(1)}_{ki} &\coloneqq \mathbb{E}_\omega \left[\varphi_k \left\langle \nabla \varphi_i, \beta \right\rangle + \alpha \left\langle \nabla \varphi_k, \nabla \varphi_i \right\rangle \right], \\ U^{(3)}_{ki} &\coloneqq \mathbb{E}_\omega [\left\langle \nabla \varphi_k, D^2 \varphi_i \beta \right\rangle]. \end{aligned}$$

Expanding $\langle g_k, A_i h \rangle$ in (6) gives

$$S_k = \frac{1}{n} \left[\sum_i K_{\rho_0}(k,i) Q_i + 2 \sum_i e_i (m_1 U^{(1)}_{ki} + m_3 U^{(3)}_{ki}) \right]. \tag{11}$$

Equations (9)–(11) prove four-moment dependence. If $m_1 = m_3 = 0$, then $Q = S = 0$ exactly, and only the even terms of (10) survive in R. □

C I.I.D. GRADIENT-FLOW ENDPOINT SELECTION

Theorem C.1 (I.i.d. gradient-flow selection)**.** *Under Assumption 4.3, full-batch gradient flow reaches a regular interpolant and*

$$\widehat{\rho}_{m,*} \xrightarrow{P} \rho_0, \qquad\qquad mc_1(\theta_{m,*}) \xrightarrow{P} \mathfrak{F}(\rho_0).$$

The first convergence includes weak convergence and every derivative moment used by $\mathfrak{F}$. The leading coefficient depends on the data, activation, and initialization law, but not on the labels.

We first formalize the topology used in Theorem C.1. For fixed data and block maps, let $\mathcal{T}_{\mathrm{DM}}$ be the finite family

$$\{\psi_i,\ \langle g_i, g_k\rangle,\ A_i[g_a, g_b],\ T_i[g_a, g_b, g_c],\ \langle g_k, A_i g_a\rangle :\ i, k, a, b, c \in [n],\}. \tag{12}$$

For measures on which these moments are finite, define

$$d_{\mathrm{DM}}(\nu, \rho) \coloneqq d_{\mathrm{BL}}(\nu, \rho) + \sum_{q \in \mathcal{T}_{\mathrm{DM}}} \left(\left| \int q d\nu - \int q d\rho \right| \wedge 1 \right).$$

On the open set where K_ρ is invertible and its top eigenvalue is simple, all entries of (7) are continuous in this metric.

Proof. Write $K_m(t) \coloneqq J(t)J(t)^\top$, let $\kappa_0 = \lambda_{\min}(K_{\rho_0}) > 0$, and let $\gamma_0 > 0$ be the population top gap. Fix an arbitrary failure probability $\alpha \in (0, 1)$. Centering makes the initial residual tight. Together with the empirical moment laws and $B_m^{16}/\sqrt{m} \to 0$ in probability, this provides deterministic $M_\alpha < \infty$ and m_α such that for every $m \geq m_\alpha$, an event $E_{m,\alpha}$ of probability at least $1 - \alpha$ satisfies

$$\|r(0)\|_2 \leq M_\alpha, \qquad \lambda_{\min}(K_m(0)) \geq \frac{3\kappa_0}{4}, \qquad \mathrm{gap}_{\mathrm{top}}(K_m(0)) \geq \frac{3\gamma_0}{4},$$

and $B_m^{16}/\sqrt{m}$ is below a deterministic threshold chosen below. The residual bound M_α is localized to $E_{m,\alpha}$, as required for a tight random variable.

Choose

$$C_\alpha > \frac{4nC_0M_\alpha}{\kappa_0}, \qquad R_{m,\alpha} \coloneqq C_\alpha \frac{B_m}{\sqrt{m}},$$

where C_0 is the fixed-data block-velocity constant. Stop the flow at the first time that either a block moves by $R_{m,\alpha}$ or

$$\|K_m(t) - K_m(0)\|_{\mathrm{op}} = \min\{\kappa_0/4, \gamma_0/8\}. \tag{13}$$

On the stopped interval the kernel floor is at least $\kappa_0/2$, and the residual flow obeys

$$\frac{d}{dt}\|r(t)\|_2^2 = -\frac{2}{n} r(t)^\top K_m(t) r(t) \leq -\frac{\kappa_0}{n}\|r(t)\|_2^2.$$

If $\tau_{m,\alpha}$ is the stopping time, then on $E_{m,\alpha}$,

$$\int_0^{t\wedge\tau_{m,\alpha}} \|r(u)\|_2 \, du \le \frac{2nM_\alpha}{\kappa_0}. \tag{14}$$

Every block velocity is bounded by $C_0 B_m \|r(t)\|_2 / \sqrt{m}$, so the integrated displacement is strictly less than $R_{m,\alpha}/2$. A mean-value estimate for the tangent features gives

$$\|K_m(t) - K_m(0)\|_{\mathrm{op}} \le C_1 B_m^2 R_{m,\alpha} = C_1 C_\alpha \frac{B_m^3}{\sqrt{m}}.$$

The chosen threshold on $B_m^{16}/\sqrt{m}$, with $B_m \ge 1$, makes this strictly smaller than (13) and also gives $R_{m,\alpha} < 1$. Both stopping inequalities improve, hence $\tau_{m,\alpha} = \infty$. Residuals decay exponentially, all velocities are integrable, every block converges, and the endpoint interpolates while retaining a positive kernel floor and top gap.

A conservative derivative-degree count yields

$$\Delta K = O_\alpha(B_m^2 R_{m,\alpha}), \qquad \Delta Q = O_\alpha(B_m^5 R_{m,\alpha}),$$
$$\Delta R = O_\alpha(B_m^6 R_{m,\alpha}), \qquad \Delta S = O_\alpha(B_m^7 R_{m,\alpha}).$$

Here is the complete count, with constants allowed to depend on fixed $(n, d, \kappa_0, \gamma_0, \alpha)$. The spectral row uses Weyl, Davis–Kahan, and the resolvent identity; all other rows follow by one product-rule expansion. We write R^{jet} for the third-jet vector to distinguish it from the displacement radius $R_{m,\alpha}$.

Object	size on the stopped tube	endpoint change
K	$O(B_m^2)$	$O_\alpha(B_m^2 R_{m,\alpha})$
(μ, e, K^{-1})	$O_\alpha(1)$	$O_\alpha(B_m^2 R_{m,\alpha})$
$h = \sum_i e_i g_i$	$O(B_m)$	$O_\alpha(B_m^3 R_{m,\alpha})$
$Q = \mu^{-1} \int A[h,h]$	$O_\alpha(B_m^3)$	$O_\alpha(B_m^5 R_{m,\alpha})$
$R^{\mathrm{jet}} = \mu^{-3/2} \int T[h,h,h]$	$O_\alpha(B_m^4)$	$O_\alpha(B_m^6 R_{m,\alpha})$
ζ	$O_\alpha(B_m^4)$	$O_\alpha(B_m^6 R_{m,\alpha})$
$S = \int \langle g, \zeta \rangle$	$O_\alpha(B_m^5)$	$O_\alpha(B_m^7 R_{m,\alpha})$

For example, changing h costs the larger of a tangent-feature variation $O(B_m R_{m,\alpha})$ and an eigenvector variation $O_\alpha(B_m^3 R_{m,\alpha})$. Substitution into Q, then into ζ and S, produces the successive exponents $5, 6, 7$ in the last column. The relaxation square is the dominant term in $\mathfrak{F}$: its first variation is

$$O_\alpha(B_m^5) O_\alpha(B_m^7 R_{m,\alpha}) = O_\alpha(B_m^{12} R_{m,\alpha}),$$

whereas the Q-square and third-jet terms cost respectively $O_\alpha(B_m^8 R_{m,\alpha})$ and $O_\alpha(B_m^6 R_{m,\alpha})$.

Since $S = O(B_m^5)$, the induced change of $\mathfrak{F}$ is at most $O_\alpha(B_m^{13}/\sqrt{m}) = o_P(1)$. The same mean-value argument applies to every test function in (12); the bounded-Lipschitz endpoint-to-start distance is bounded by the maximal block displacement. Empirical measures converge weakly at initialization, and the integrability assumptions give laws of large numbers for all derivative moments. Therefore, for every fixed $\delta > 0$,

$$\Pr\left(d_{\mathrm{DM}}(\widehat{\rho}_{m,*}, \rho_0) > \delta, E_{m,\alpha}\right) \longrightarrow 0.$$

The full failure probability is at most $\alpha + o(1)$. Sending the arbitrary α to zero proves convergence in probability. Continuity of $\mathfrak{F}$ and Theorem 4.1 prove the coefficient limit. Centered block outputs place the initial residual on the tight scale used by the lazy endpoint argument; an uncentered mean can instead produce a $\Theta_P(\sqrt{m})$ residual. □

C.1 AUTOMATIC REGULARITY FOR SOFTPLUS

Corollary C.2 (Concrete Softplus model)**.** *Let the inputs be distinct points in the Euclidean unit ball. For Softplus, i.i.d. standard-Gaussian hidden weights and biases, and an independent nondegenerate centered subgaussian readout, all conditions of Theorem C.1 hold. Moreover,*

$$\max_j \|\theta_{j,*} - \theta_{j,0}\|_2 = O_P\left(\frac{\sqrt{\log m}}{\sqrt{m}}\right).$$

Proof. The readout-coordinate feature kernel is

$$K_{ik}^{\text{feat}} = \mathbb{E}_{w,b}[s(w^\top x_i + b)s(w^\top x_k + b)].$$

If a nonzero coefficient vector a had zero quadratic form, Gaussian full support and continuity would imply $\sum_i a_i s(w^\top x_i + b) = 0$ for all (w, b). Differentiating twice in b and Fourier transforming gives

$$\widehat{s''}(t) \sum_i a_i e^{itw^\top x_i} = 0.$$

For Softplus, s'' is the logistic density and $\widehat{s''}(t) = \pi t/\sinh(\pi t)$, continuously extended at zero, never vanishes on the real line. Varying w and using uniqueness of the Fourier transform of the finite signed measure $\sum_i a_i \delta_{x_i}$ forces $a = 0$. Hence K^{feat} is strictly positive definite.

The full tangent kernel is

$$K_{ik} = K_{ik}^{\text{feat}} + m_2(1 + x_i^\top x_k)\mathbb{E}[s'(w^\top x_i + b)s'(w^\top x_k + b)].$$

The second term is a feature Gram matrix, so strict positivity is preserved. Every entry is positive, and Perron–Frobenius gives a simple top eigenvalue. Centering makes the initial residual tight. On the unit enlargement, the largest readout and preactivation are $O_P(\sqrt{\log m})$. Because the block output is their product up to an additive constant, the common map-and-jet envelope in Assumption 4.3 is $B_m = O_P(\log m)$, which still gives $B_m^{16}/\sqrt{m} \to 0$ in probability.

After the preceding bootstrap has established the infinite-time kernel floor, only first parameter derivatives enter the velocity bound. Their sharper envelope is $G_m = O_P(\sqrt{\log m})$. Reusing (14) with G_m therefore gives

$$\max_j \|\theta_{j,*} - \theta_{j,0}\|_2 = O_P\left(\frac{\sqrt{\log m}}{\sqrt{m}}\right).$$

The displayed rate is pointwise in the fixed dataset because its smallest kernel eigenvalue may depend on that dataset. □

D TANGENT FUNCTIONALS AND THEIR SELECTED LAWS

D.1 EXACT TANGENT CONTRACTION

Proof. At interpolation, simple-eigenvalue differentiation gives $\nabla \mathsf{S} = z = (J^\top q + 2B_v^\top u)/n$. The tangent projector is $P_T = I - J^\top K^{-1} J$. In contrast, the gradient of the Gauss–Newton top eigenvalue is $2B_v^\top u/n$; its tangent projection agrees with $P_T z$, since $P_T J^\top q = 0$. The block scaling in Theorem 4.1 gives $z_j = \zeta_{\widehat{\rho}_m}(\theta_j)/m$ and $Jz = S(\widehat{\rho}_m)/\sqrt{m}$. Consequently,

$$m\,\|P_T z\|^2 = m\left(\|z\|^2 - (Jz)^\top K^{-1}(Jz)\right)$$

$$= \int \|\zeta_{\widehat{\rho}_m}\|^2 \, \mathrm{d}\widehat{\rho}_m - S(\widehat{\rho}_m)^\top K_{\widehat{\rho}_m}^{-1} S(\widehat{\rho}_m).$$

The identities $\zeta = \chi + \sum_i Q_i g_i/n$ and $S = U + KQ/n$ cancel every Q term after expansion. For any law ρ, completing the square yields

$$\begin{aligned}\int \|\chi - \textstyle\sum_i c_i g_i\|^2 \, \mathrm{d}\rho &= \int \|\chi\|^2 \, \mathrm{d}\rho - 2c^\top U + c^\top K c \\ &= \mathcal{D}(\rho) + (c - K^{-1}U)^\top K (c - K^{-1}U).\end{aligned}$$

Thus $\mathcal{D}$ is nonnegative, with equality precisely when χ belongs to the span of the sample features in $L^2(\rho)$. Changing e to $-e$ leaves χ unchanged, so the functional is orientation invariant. □

D.2 Transport of the tangent moments

Proof. Enlarge the finite derivative-test family in (12) by $\langle A_i g_a, A_j g_b \rangle$ for $i, j, a, b \in [n]$. Their assumed integrability gives the initialization law of large numbers. The kernel floor and simple top gap make K^{-1} and a locally oriented e continuous; $\mathcal{D}$ is a continuous expression in these finitely many moments.

For the endpoint comparison, fix failure probability $\alpha > 0$ and use the stopped tube from the proof of Theorem C.1. Its maximal block radius is $R_{m,\alpha} = O_\alpha(B_m/\sqrt{m})$. The kernel changes by $O_\alpha(B_m^2 R_{m,\alpha})$, as do its inverse and top eigenvector up to fixed floor and gap factors. Product differentiation then gives

$$\begin{aligned} \|h\| &= O(B_m), & \Delta h &= O_\alpha(B_m^3 R_{m,\alpha}), \\ \|\chi\| &= O(B_m^2), & \Delta \chi &= O_\alpha(B_m^4 R_{m,\alpha}), \\ \|U\| &= O(B_m^3), & \Delta U &= O_\alpha(B_m^5 R_{m,\alpha}). \end{aligned}$$

Here changes include both moving the blocks and changing the empirical eigenvector. Expanding the two terms of $\mathcal{D}$ gives

$$\begin{aligned} \Delta \int \|\chi\|^2 \, \mathrm{d}\widehat{\rho}_m &= O_\alpha(B_m^6 R_{m,\alpha}), \\ \Delta(U^\top K^{-1} U) &= O_\alpha(B_m^8 R_{m,\alpha}), \\ \mathcal{D}(\widehat{\rho}_{m,*}) - \mathcal{D}(\widehat{\rho}_{m,0}) &= O_\alpha(B_m^9/\sqrt{m}) = o_P(1). \end{aligned}$$

The last convergence follows from $B_m \geq 1$ and the assumed $B_m^{16}/\sqrt{m} \to 0$ condition. Combining initialization concentration with endpoint transport and then sending α to zero proves the joint limit. Convergence of the derivative moments supplies the required strengthening of weak convergence. □

D.3 Four readout moments and nonzero tangent motion

For independent readout and hidden parameters, define $\alpha \coloneqq \sum_i e_i \varphi_i$, $\beta \coloneqq \nabla \alpha$, and $\mathsf{H} \coloneqq \nabla^2 \alpha$. Expectations in the next display are over the hidden law. Put

$$u_k \coloneqq \mathbb{E}[\varphi_k \|\beta\|^2 + \alpha \langle \nabla \varphi_k, \beta \rangle], \qquad w_k \coloneqq \mathbb{E} \langle \nabla \varphi_k, \mathsf{H}\beta \rangle .$$

Proof. Block multiplication gives

$$g_i = (\varphi_i, A\nabla\varphi_i), \qquad h = (\alpha, A\beta), \qquad \chi = \frac{2}{n}(A\|\beta\|^2, \alpha\beta + A^2\mathsf{H}\beta),$$
$$U = \frac{2}{n}(m_1 u + m_3 w).$$

Expanding the squared norm and integrating over A proves the exact formula

$$\begin{aligned}\mathcal{D}(\rho_0) = \frac{4}{n^2}\Big\{ &\mathbb{E}\big[\alpha^2 \|\beta\|^2 + m_2(\|\beta\|^4 + 2\alpha\beta^\top \mathsf{H}\beta) + m_4 \|\mathsf{H}\beta\|^2\big] \\ &- (m_1 u + m_3 w)^\top K^{-1}(m_1 u + m_3 w)\Big\}. \end{aligned} \tag{15}$$

The kernel and e depend only on m_2. Hence no higher readout moment enters. When $m_1 = m_3 = 0$, the subtraction vanishes, proving the positive square representation in the theorem. If $m_2 > 0$, its first summand implies $\mathcal{D} = 0$ only if $\beta = 0$ almost surely; that condition also makes both summands zero. For common m_2 and balanced laws, subtracting (15) shows

$$\mathcal{D}'' - \mathcal{D}' = \frac{4}{n^2}(m_4'' - m_4')\mathbb{E}\left\|\mathsf{H}\beta\right\|^2.$$

Thus a fourth-moment intervention preserves the population kernel while potentially changing both the drift and the flip coefficient. □

E Proof of the neural trajectory law

We first give a deterministic result for possibly different block maps. Fix a block dimension and constants $L \geq 1$, $M \geq 0$, $\lambda_- > 0$, $D_- > 0$, $A_0 < \infty$. At an interpolant of $f_m = \epsilon \sum_j \psi_j$, put

$$\begin{aligned}
h_j &\coloneqq \nabla\psi_j(\theta_{j,*}), & A_j &\coloneqq D^2\psi_j(\theta_{j,*}),\\
g_* &\coloneqq \nabla f_m(\theta_*), & H_* &\coloneqq D^2 f_m(\theta_*),\\
\lambda_* &\coloneqq m^{-1}\sum_j \left\|h_j\right\|^2, & \kappa_* &\coloneqq m^{-1}\sum_j h_j^\top A_j h_j,\\
v_* &\coloneqq g_*/\sqrt{\lambda_*}, & P_* &\coloneqq I - v_* v_*^\top,\\
B_* &\coloneqq P_* H_* g_*/\epsilon, & D_m &\coloneqq 4\left\|B_*\right\|^2.
\end{aligned}$$

In particular, $D_m = 4(m^{-1}\sum_j \left\|A_j h_j\right\|^2 - \kappa_*^2/\lambda_*)$.

Theorem E.1 (Deterministic paired-GD approximation)**.** *Suppose each ψ_j is C^4 with derivative operator norms of orders one through four at most L on the unit ball about $\theta_{j,*}$, and*

$$\lambda_* \geq \lambda_-, \qquad |\kappa_*| \leq M\epsilon, \qquad D_m \geq D_-.$$

Let $|a_0|, |d_0| \leq A_0$, $\eta_* \coloneqq 2/\lambda_*$, and run GD with $\eta_m \coloneqq (2+\epsilon d_0)/\lambda_*$ from $\theta_* + a_0 v_*$. The functions (a_m, d_m) solve

$$a_m' = 2d_m a_m, \qquad d_m' = -\eta_*^2 D_m a_m^2, \qquad (a_m(0), d_m(0)) = (a_0, d_0). \tag{16}$$

For every fixed $T < \infty$, there are C_T, m_T depending only on the displayed constants and T such that, for $m \geq m_T$, all iterates through $2\lfloor T/\epsilon\rfloor + 1$ stay within the derivative neighborhoods, and

$$\max_{k \leq T/\epsilon}\left\{\left|\frac{r(\theta^{(2k)})}{\sqrt{\lambda_*}} - a_m(\epsilon k)\right| + \left|\frac{\eta_m\left\|\nabla f_m(\theta^{(2k)})\right\|^2 - 2}{\epsilon} - d_m(\epsilon k)\right|\right\} \leq C_T\epsilon,$$

$$\max_{\ell \leq 2\lfloor T/\epsilon\rfloor + 1}\max_j \left\|\theta_j^{(\ell)} - \theta_{j,*}\right\| \leq C_T\epsilon. \tag{17}$$

Uniformly on these iterates, $|\mathsf{S} - \left\|\nabla f_m\right\|^2| \leq C_T\epsilon^2$. Odd residuals equal minus the preceding even residual plus $O_T(\epsilon)$, and odd and even Gauss–Newton sharpness differ by $O_T(\epsilon^2)$. Furthermore, with $w_m(t) \coloneqq 2(d_m(t) - d_0)B_*/(\eta_* D_m)$,

$$\max_{k \leq T/\epsilon}\left\|\theta^{(2k)} - \theta_* - a_m(\epsilon k)v_* - w_m(\epsilon k)\right\| \leq C_T\epsilon, \qquad \left\|w_m(t)\right\| = \frac{|d_m(t) - d_0|}{\eta_*\sqrt{D_m}}. \tag{18}$$

Proof. Write $g \coloneqq \nabla f_m$, $H \coloneqq D^2 f_m$, $\lambda \coloneqq \left\|g\right\|^2$, and $b \coloneqq Hg/\epsilon$. Initially stop the trajectory when $|r|$ or $|d| = |(\eta_m\lambda - 2)/\epsilon|$ exceeds a fixed K, or a block moves more than $R\epsilon$. We choose K, R below. Within this region, blockwise differentiation gives

$$\begin{aligned}
g_j &= \epsilon\nabla\psi_j, & H_{jj} &= \epsilon D^2\psi_j, & b_j &= \epsilon D^2\psi_j\nabla\psi_j,\\
\left\|g\right\| + \left\|b\right\| &= O(1), & \left\|H\right\|_{\mathrm{op}} &= O(\epsilon), & \max_j(\left\|g_j\right\| + \left\|b_j\right\|) &= O(\epsilon).
\end{aligned}$$

Bounded third derivatives and maximal block motion $R\epsilon$ give

$$\left\|g - g_*\right\| + \left\|b - H_* g_*/\epsilon\right\| = O_R(\epsilon),$$

$$m^{-1}\sum_j (\nabla\psi_j)^\top D^2\psi_j \nabla\psi_j = \kappa_* + O_R(\epsilon) = O_R(\epsilon),$$

$$g^\top H g = O_R(\epsilon^2), \qquad \|b\|^2 = D_m/4 + O_R(\epsilon).$$

For the last identity, decompose $H_* g_*/\epsilon = B_* + \kappa_* g_*/\lambda_*$; its squared normal component is $\kappa_*^2/\lambda_* = O(\epsilon^2)$.

The needed Taylor estimates concern dense directions, not the full third-derivative tensor. If $\max_j \|z_j\| \le C\epsilon$, then $\|z\| = O(1)$ and

$$|D^3 f_m[z,z,z]| \le \epsilon L \sum_j \|z_j\|^3 = O(\epsilon^2),$$

$$\left\|D^3 f_m[z,z,\cdot]\right\| \le \epsilon L \left(\sum_j \|z_j\|^4\right)^{1/2} = O(\epsilon^2),$$

$$D^2\lambda = 2H^2 + 2D^3 f_m[g,\cdot,\cdot], \qquad \left\|D^2\lambda\right\|_{\mathrm{op}} = O(\epsilon^2). \tag{19}$$

The vector remainder in the second line has each block $O(\epsilon^3)$. The final bound follows because the jth block of the contraction is $\epsilon D^3\psi_j[g_j,\cdot,\cdot] = O(\epsilon^2)$.

At an even iterate, one GD step has increment $-\eta_m r g$. Taylor expansion along that dense increment yields

$$\begin{aligned} r_+ &= (1-\eta_m\lambda)r + \frac{\eta_m^2 r^2}{2} g^\top H g + O(\epsilon^2) \\ &= -r - \epsilon d r + O_R(\epsilon^2), \\ g_+ &= g - \eta_m r H g + E_g, \qquad \|E_g\| = O(\epsilon^2), \qquad \max_j \|(E_g)_j\| = O(\epsilon^3). \end{aligned}$$

Substitute both expressions into the exact paired update. The two order-one normal increments cancel:

$$\begin{aligned} \Delta\theta &= -\eta_m r g - \eta_m r_+ g_+ = \eta_m \epsilon d r g - \eta_m^2 r^2 H g + E_\theta, \\ \|E_\theta\| &= O_R(\epsilon^2), \qquad \max_j \|(E_\theta)_j\| = O_R(\epsilon^3). \end{aligned} \tag{20}$$

Thus $\|\Delta\theta\| = O(\epsilon)$ and $\max_j \|\Delta\theta_j\| = O(\epsilon^2)$. Taylor expansion across this smaller paired increment gives

$$\begin{aligned} \Delta r &= g^\top \Delta\theta + O(\epsilon \|\Delta\theta\|^2) \\ &= \eta_m \epsilon d r \lambda - \eta_m^2 r^2 g^\top H g + O_R(\epsilon^2) = 2\epsilon d r + O_R(\epsilon^2), \\ \Delta\lambda &= 2(Hg)^\top \Delta\theta + O(\epsilon^2 \|\Delta\theta\|^2) \\ &= 2\eta_m \epsilon d r g^\top H g - 2\eta_m^2 r^2 \|Hg\|^2 + O_R(\epsilon^3) \\ &= -2\eta_m^2 \epsilon^2 r^2 \|b\|^2 + O_R(\epsilon^3). \end{aligned}$$

With $\widehat{a} = r/\sqrt{\lambda_*}$ and $\widehat{d} = (\eta_m\lambda - 2)/\epsilon$, it follows that

$$\begin{aligned} \Delta\widehat{a} &= 2\epsilon \widehat{d}\widehat{a} + O_R(\epsilon^2), \\ \Delta\widehat{d} &= -2\eta_m^3 \lambda_* \epsilon \widehat{a}^2 \|b\|^2 + O_R(\epsilon^2) = -\eta_*^2 D_m \epsilon \widehat{a}^2 + O_R(\epsilon^2). \end{aligned} \tag{21}$$

The last equality uses $\eta_*\lambda_* = 2$, $\eta_m = \eta_* + O(\epsilon)$, and $\|b\|^2 = D_m/4 + O_R(\epsilon)$.

At the initial normal perturbation, $v_*^\top H_* v_* = \epsilon\kappa_*/\lambda_* = O(\epsilon^2)$. Applying (19) gives

$$r(\theta_* + a_0 v_*) = \sqrt{\lambda_*} a_0 + O(\epsilon^2), \qquad \lambda(\theta_* + a_0 v_*) = \lambda_* + O(\epsilon^2).$$

Hence the initial state error is $O(\epsilon)$. Direct differentiation of (16) gives the invariant

$$d_m(t)^2 + \frac{\eta_*^2 D_m}{2} a_m(t)^2 = d_0^2 + \frac{\eta_*^2 D_m}{2} a_0^2.$$

Since $\lambda_* \in [\lambda_-, L^2]$ and $D_m \in [D_-, 4L^4]$, it bounds the ODE uniformly. Choose K above those state bounds and the corresponding residual bounds by a fixed margin. On this compact state region the vector field is uniformly Lipschitz. Comparing (21) to the ODE Euler expansion gives, for a constant C after R is fixed,

$$e_{k+1} \le (1 + C\epsilon)e_k + C\epsilon^2, \qquad e_0 \le C\epsilon,$$

$$e_k \le (1 + C\epsilon)^k e_0 + C\epsilon^2 \sum_{j=0}^{k-1} (1 + C\epsilon)^j \le C_T\epsilon \qquad (k \le T/\epsilon).$$

For a noncircular neighborhood bootstrap, retain the separate orders in the block bound from (20):

$$\max_j \|\Delta\theta_j\| \le C_K\epsilon^2 + C_K(1 + M + R)\epsilon^3.$$

The leading constant uses only the state bound and unit-neighborhood derivative bounds. The initial and odd-step block displacements are at most $A_0 L\epsilon/\sqrt{\lambda_-}$ and $C_K\epsilon$, respectively. First choose $R > A_0 L/\sqrt{\lambda_-} + C_K T + C_K + 2$. Then choose m sufficiently large that the accumulated higher-order term and state error fit within these margins and $(R + C_K)\epsilon < 1$. This also contains all Taylor segments. Both stopping conditions remain inactive, proving (17) and the block bound.

To compare actual and Gauss–Newton sharpness, use $D^2 L_m = gg^\top + rH$ and split along $u = g/\sqrt{\lambda}$ and $u^\perp$. The matrix blocks have the form

$$\begin{pmatrix} \lambda + e_{00} & b_{01}^\top \\ b_{01} & C_{11} \end{pmatrix}, \qquad e_{00} = \frac{rg^\top Hg}{\lambda} = O_R(\epsilon^2), \qquad \|b_{01}\| + \|C_{11}\| = O(\epsilon).$$

The variational bound $|\mathsf{S} - \lambda| \le \|rH\| = O(\epsilon)$ and the positive lower bound on λ imply $\left\|(\mathsf{S}I - C_{11})^{-1}\right\| = O(1)$. The Schur complement therefore yields

$$\mathsf{S} - \lambda - e_{00} = b_{01}^\top (\mathsf{S}I - C_{11})^{-1} b_{01} = O(\epsilon^2).$$

This proves the sharper $O_R(\epsilon^2)$ error at both parities. The first-step formula gives $\lambda_+ - \lambda = -2\eta_m r g^\top Hg + O(\epsilon^2) = O_R(\epsilon^2)$, while the residual formula gives the odd-step claim. Squaring the bounded residual proves the loss approximation.

Finally, replace g by $\sqrt{\lambda_*}v_* + O_R(\epsilon)$ and Hg/ϵ by $B_* + O_R(\epsilon)$ in (20), and use the scalar-state error already proved:

$$\Delta\theta = \epsilon \left[2d_m a_m v_* - \eta_*^2 \lambda_* a_m^2 B_*\right] + O_T(\epsilon^2).$$

The bracket equals $(a_m v_* + w_m)'$, since $w_m' = -2\eta_* a_m^2 B_*$ and $\eta_*\lambda_* = 2$. Summing the vector local errors over at most T/ϵ pairs proves (18). Orthogonality and $\|B_*\| = \sqrt{D_m}/2$ give its displayed tangent-displacement norm. □

E.1 Compact i.i.d. selection and quantitative transfer

Proof. For one residual, the positive scalar kernel has gap λ_∞ from zero; compact support gives a bounded derivative envelope. With centering, these are the hypotheses of Theorem C.1. Its stopping argument gives maximal gradient-flow block displacement $O_P(\epsilon)$: the initial residual is tight, the scalar kernel stays bounded below, and the integrated residual bounds each block velocity. The two-unit initialization enlargement therefore contains the unit neighborhoods of the regular selected interpolant with probability tending to one. The bounded, locally Lipschitz moments $\|h\|^2$, $h^\top Ah$, and $\|Ah\|^2$ have i.i.d. fluctuations $O_P(\epsilon)$ by Chebyshev and endpoint changes $O_P(\epsilon)$, giving

$$\lambda_* = \lambda_\infty + O_P(\epsilon), \qquad \kappa_* = O_P(\epsilon), \qquad D_m = D_\infty + O_P(\epsilon).$$

Thus Theorem E.1 applies on localized events. The invariant bounds its ODE on each fixed interval, where replacing $\eta_*^2 D_m$ by $4D_\infty/\lambda_\infty^2$ changes the solution by $O_P(\epsilon)$ through Gronwall. Combine this with (17) and the loss-sharpness estimate. □

E.2 Explicit trajectories

Put $c_m := \eta_*^2 D_m/2$ and $R_m := \sqrt{d_0^2 + c_m a_0^2}$. If $a_0 = 0$, the ODE is stationary. Otherwise $u_0 := \operatorname{artanh}(d_0/R_m)$ is finite and direct substitution in (16) gives

$$d_m(t) = -R_m \tanh(2R_m t - u_0), \qquad a_m(t) = \operatorname{sgn}(a_0)\frac{R_m}{\sqrt{c_m}} \operatorname{sech}(2R_m t - u_0). \tag{22}$$

E.3 An exact neural witness and the projection comparison

Proof. For $\psi(A,\beta) = As(\beta^\top \widetilde{x})$ with $\widetilde{x} = 1$, put $(s,p,r,t) \coloneqq (s,s',s'',s''')$ at the initialized bias. Direct differentiation gives

$$h = (s, Ap), \qquad Ah = (Ap^2, sp + A^2pr), \quad h^\top Ah = 2Asp^2 + A^3rp^2,$$
$$D^3\psi[h,h,h] = 3A^2srp^2 + A^4tp^3.$$

Symmetric readouts cancel $\mathbb{E}\psi$ and $\mathbb{E}[h^\top Ah]$. The two laws have $m_2 = 1$ and $m_4 = 1, 16$. At bias $\log 3$, Softplus has $(s,p,r,t) = (\ell, 3/4, 3/16, -3/32)$. The scalar balanced specialization of (7) and the squared norm of Ah give

$$\lambda_\infty = s^2 + p^2, \qquad \mathfrak{F} = -\frac{4}{\lambda_\infty^2}\left(srp^2 + \frac{m_4}{3}tp^3\right), \qquad D_\infty = 4\left[p^4 + p^2(s^2 + 2sr + m_4r^2)\right].$$

Substitution proves the displayed constants and their signs. The law has compact support and Softplus is smooth, so the derivative bounds hold on its two-unit enlargement. Positivity of s, p, r gives $\lambda_\infty, D_\infty > 0$. The selected flip law follows from Theorem C.1. On arbitrary inputs, the initialized hidden preactivation is still $\log 3$, and $\mathbb{E}A^2 = 1$ gives $K_\infty(x,x') = \ell^2 + (9/16)\langle \widetilde{x}, \widetilde{x}'\rangle$. Thus the two laws share the same population kernel; their finite i.i.d. empirical kernels concentrate around that common limit. □

E.4 The fixed-normal projected map

Proof. First use the deterministic assumptions and let $\phi(p) \coloneqq f_m(\theta_* + pv_*) - y$. The exact fixed-normal projected update is $p_{\ell+1} = p_\ell - \eta_m \phi(p_\ell)\phi'(p_\ell)$. On every fixed bounded p interval,

$$\phi''(0) = \epsilon\kappa_*/\lambda_* = O(\epsilon^2), \qquad |\phi'''(p)| \le \epsilon L \sum_j \|v_{j,*}\|^3 = O(\epsilon^2),$$
$$\phi'(p) = \sqrt{\lambda_*} + O(\epsilon^2), \qquad \phi(p) = \sqrt{\lambda_*}p + O(\epsilon^2).$$

Consequently,

$$p_{\ell+1} = -(1 + \epsilon d_0)p_\ell + O(\epsilon^2), \qquad p_{2k+2} - p_{2k} = 2\epsilon d_0 p_{2k} + O(\epsilon^2).$$

A stopped scalar Euler comparison, with the amplitude bound chosen above the finite-T exponential solution, yields $p_{2k} = a_0 \exp(2d_0\epsilon k) + O_T(\epsilon)$. This is the exact projected map and includes the higher-order terms.

For a scalar residual the exact coefficient formula reduces to

$$c_1 = \frac{8q^2}{\lambda_*} - \frac{4t}{3\sqrt{\lambda_*}}, \qquad q = D^2 f_m[v_*, v_*], \quad t = D^3 f_m[v_*, v_*, v_*].$$

The preceding dense estimates give $q, t = O(\epsilon^2)$, hence $c_1 = O(\epsilon^2)$. The specified common detuning therefore has $d_{0,m} = |c_1|\,(2a_0)^2/\epsilon = O(\epsilon)$. The deterministic theorem is uniform over bounded width-dependent d_0, so its finite-T ODE changes by only $O_T(\epsilon)$ when $d_{0,m}$ is replaced by zero. In (22), $d_0 = 0$ gives $a_m(t) = a_0 \operatorname{sech}(\eta_*\sqrt{2D_m}|a_0|t)$, whereas the projected amplitude is $a_0 + O_T(\epsilon)$. The compact i.i.d. transfer replaces (η_*, D_m) by $(2/\lambda_\infty, D_\infty)$ and deterministic errors by $O_P(\epsilon)$, proving the corollary. □

F One trajectory from initialization to its own interpolant

We first state the trajectory law for a perturbation of a selected interpolant. A weighted expansion then transfers this law to random initialization, and a summable tail proves convergence to the GD trajectory's own endpoint.

F.1 Dynamics near a selected interpolant

Theorem F.1 (Selected neural trajectory law)**.** *Let ψ be C^4 on a fixed-dimensional block space. Initialize i.i.d. from a compactly supported law ρ_0, and suppose derivatives through order four are bounded on its two-unit enlargement. For $h \coloneqq \nabla\psi$ and $A \coloneqq D^2\psi$, assume*

$$\mathbb{E}\psi = 0, \qquad \lambda_\infty \coloneqq \mathbb{E}\,\|h\|^2 > 0, \qquad \mathbb{E}[h^\top Ah] = 0, \qquad D_\infty \coloneqq 4\mathbb{E}\,\|Ah\|^2 > 0.$$

For every fixed finite label y, gradient flow reaches a regular interpolant θ_* with probability tending to one. Put $\lambda_* \coloneqq \|\nabla f_m(\theta_*)\|^2$ and $v_* \coloneqq \nabla f_m(\theta_*)/\sqrt{\lambda_*}$. Initialize unprojected GD at $\theta^{(0)} \coloneqq \theta_* + a_0 v_*$ with step $\eta_m \coloneqq (2+\epsilon d_0)/\lambda_*$, for fixed $a_0, d_0 \in \mathbb{R}$. The functions (a, d) solve

$$a' = 2da, \qquad d' = -\left(\frac{2}{\lambda_\infty}\right)^2 D_\infty a^2, \qquad (a(0), d(0)) = (a_0, d_0).$$

For every fixed $T < \infty$, uniformly over integer $0 \le k \le T/\epsilon$,

$$\frac{f_m(\theta^{(2k)}) - y}{\sqrt{\lambda_*}} = a(\epsilon k) + O_P(\epsilon), \qquad \frac{\eta_m \mathsf{S}(\theta^{(2k)}) - 2}{\epsilon} = d(\epsilon k) + O_P(\epsilon),$$

where S is the largest actual loss-Hessian eigenvalue. Moreover, every block moves $O_P(\epsilon)$ from its selected endpoint throughout these $O(\sqrt{m})$ steps. The loss is $\lambda_* a(\epsilon k)^2/2 + O_P(\epsilon)$ at both parities.

The time variable counts pairs: $t = \epsilon k$ corresponds to $2t\sqrt{m}$ individual GD steps. The theorem is uniform on each fixed slow-time interval, and Appendix F adds a summable tail for fixed-width convergence. Its proof and deterministic counterpart appear in Appendix E; the weighted estimates below give the finite-moment extension used in the main theorem.

F.2 Derivative averages and the initialization expansion

Write $\epsilon \coloneqq m^{-1/2}$, $r \coloneqq f_m - y$, $g \coloneqq \nabla f_m$, $H \coloneqq D^2 f_m$, $\lambda \coloneqq \|g\|^2$, and $\kappa \coloneqq m^{-1}\sum_j h_j^\top A_j h_j$, where $h_j \coloneqq \nabla\psi(\theta_j)$ and $A_j \coloneqq D^2\psi(\theta_j)$. At initialization define

$$\lambda_0 = \|g_0\|^2, \qquad v_0 = g_0/\sqrt{\lambda_0}, \quad B_0 = (I - v_0 v_0^\top) H_0 g_0/\epsilon,$$
$$D_0 = 4\|B_0\|^2 = 4\left(\frac{1}{m}\sum_j \|A_j h_j\|^2 - \frac{\kappa_0^2}{\lambda_0}\right), \quad a_0 \coloneqq r_0/\sqrt{\lambda_0}.$$

For a fixed step η, let $d \coloneqq (\eta\lambda - 2)/\epsilon$ and $\eta_0 \coloneqq 2/\lambda_0$.

Lemma F.2 (Weighted initialization transfer)**.** *Suppose the block maps are C^4 on unit balls around initialization and their first four derivative norms there are bounded by $u_j \ge 1$, with $m^{-1}\sum_j u_j^8 \le C_u$. Suppose also $\lambda_0 \ge \lambda_- > 0$, $|\kappa_0| \le M\epsilon$, $D_0 \ge D_- > 0$, and $|a_0|, |d_0| \le A$, where $\eta = (2+\epsilon d_0)/\lambda_0$. Start actual GD at the unmodified initialization. For every fixed T, the conclusions of Theorem E.1 hold with initial amplitude a_0, reference kernel λ_0, and coefficient D_0. The parameter formula becomes*

$$\theta^{(2k)} = \theta^0 + [a(\epsilon k) - a_0]v_0 + \frac{2[d(\epsilon k) - d_0]}{\eta_0 D_0} B_0 + O_T(\epsilon), \tag{23}$$

and block j moves at most $C_T \epsilon u_j^2$. The constants depend only on the displayed bounds and T.

Proof. Stop inside a bounded (r, d) region and block neighborhoods of radii $R\epsilon u_j^2$. The moment bound implies $\epsilon \max_j u_j^2 \le C_u^{1/4} m^{-1/4} \to 0$, so these lie inside the unit derivative balls. With $b = Hg/\epsilon$, block differentiation gives

$$\|g_j\| \le \epsilon u_j, \qquad \|H_{jj}\| \le \epsilon u_j, \qquad \|(Hg)_j\| \le \epsilon^2 u_j^2,$$
$$\|g - g_0\| \le C_R\epsilon(\overline{u^6})^{1/2}, \qquad \|b - H_0 g_0/\epsilon\| \le C_R\epsilon(\overline{u^8})^{1/2}.$$

Here an overline denotes the block average. The derivative of $h^\top A h$ is bounded by $3u_j^3$, giving $\kappa = \kappa_0 + O(R\epsilon\overline{u^5}) = O_R(\epsilon)$. Thus $g^\top H g = O_R(\epsilon^2)$ and $\|b\|^2 = D_0/4 + O_R(\epsilon)$.

For the first increment $z = -\eta r g$, Taylor's theorem gives

$$r_+ = (1 - \eta\lambda)r + E_r, \qquad |E_r| \le C_R\epsilon^2 r^2,$$
$$g_+ = g - \eta r H g + E_g, \qquad \|E_g\| \le C\epsilon^2 r^2, \qquad \|(E_g)_j\| \le C\epsilon^3 u_j^3 r^2.$$

For example the scalar cubic remainder is at most $C\epsilon \sum_j u_j(\epsilon u_j|r|)^3 = C\epsilon^2\overline{u^4}|r|^3$. Substitution into the exact second step yields

$$\Delta\theta = \eta\epsilon dr g - \eta^2 r^2 Hg + E_\theta, \qquad \|E_\theta\| \le C_R\epsilon^2 r^2, \qquad \|(E_\theta)_j\| \le C_R\epsilon^3 u_j^3 r^2. \tag{24}$$

In particular, its norm is $O(\epsilon(|r|+r^2))$ and its block norms are $O(\epsilon^2 u_j^2(|r|+r^2))$. Taylor expansion across the pair then gives the residual-relative estimates

$$\begin{aligned} r_{++} &= (1+2\epsilon d)r + E_a, & |E_a| &\le C_R\epsilon^2|r|, \\ d_{++} - d &= -2\eta^3\epsilon r^2\|b\|^2 + E_d, & |E_d| &\le C_R\epsilon^2(|r|+r^2). \end{aligned} \tag{25}$$

The term $\epsilon^2 d^2 r$ explains the linear residual factor in E_a. For the second line, $D^2\lambda$ has block norm at most $C\epsilon^2 u_j^2$, so its paired remainder is $O(\epsilon^4\overline{u^6}r^2)$. The inner product of Hg with E_θ is $O(\epsilon^3\overline{u^5}r^2)$. The potentially positive linear error in Δd is $2\eta^2\epsilon dr\kappa = O_R(\epsilon^2|r|)$.

Replacing $\|b\|^2$ by $D_0/4 + O_R(\epsilon)$ gives the Euler system $\Delta a = 2\epsilon da + O(\epsilon^2)$, $\Delta d = -\eta_0^2 D_0\epsilon a^2 + O(\epsilon^2)$. Both initial coordinates match exactly. Its invariant $d^2 + \eta_0^2 D_0 a^2/2$ bounds the limiting ODE. Finite-time discrete Gronwall gives the $O_T(\epsilon)$ scalar error. Summing paired block motion, together with one intervening odd step, gives $O_T(\epsilon u_j^2)$ displacement. Choose the scalar stopping bounds first, then R from these leading bounds, then m large enough to absorb the R-dependent higher-order errors. No first exit occurs.

For actual sharpness, $D^2L = gg^\top + rH$. Although $\|H\| \le \epsilon \max u_j$ need not be $O(\epsilon)$, it tends to zero. For $v = g/\sqrt{\lambda}$, $|v^\top Hv| = O(\epsilon^2)$ and $\|Hv\|^2 = O(\epsilon^2)$. The complementary resolvent in the rank-one Schur complement is bounded by the kernel floor. Its coupling norm is at most $|r|\,\|Hv\|$, proving $|\mathsf{S} - \lambda| = O_T(\epsilon^2)$ at both parities. Finally replace g by g_0, b by B_0, and η by η_0 in (24); the resulting error is $O(\epsilon^2)$ in norm. The leading vector is the derivative of (23), which proves the parameter conclusion by summation. □

F.3 A SUMMABLE TAIL AT FIXED WIDTH

Lemma F.3 (Tail contraction and interpolation)**.** *Under the preceding stopped-region estimates, suppose an even iterate has $-K_0 \le d \le -2\beta < 0$ and $|r| \le \rho$. For sufficiently small fixed $\rho > 0$ and sufficiently large m, GD converges to a regular interpolant, and its subsequent paired residuals satisfy $|r_{k+j}| \le (1-\beta\epsilon)^j|r_k|$. The total additional block-i displacement is at most $C\epsilon u_i^2\rho/\beta$.*

Proof. Stop if $|r| > \rho$, $d \notin [-K_0 - 1, -\beta]$, or the additional weighted block excursion exceeds ϵu_i^2. In this region (25) has residual multiplier in $[0, 1-\beta\epsilon]$ at large width. Hence every stopped prefix satisfies

$$\sum_j |r_j| \le \frac{\rho}{\beta\epsilon}, \qquad \sum_j r_j^2 \le \frac{\rho^2}{\beta\epsilon}.$$

The nonpositive leading sharpness term and its summable error give

$$d_k - C\rho^2/\beta - C_R\epsilon\rho/\beta \le d_{k+j} \le d_k + C_R\epsilon\rho/\beta.$$

Choose ρ so the first fixed loss is below $1/4$, then m so the epsilon terms are below $\min(\beta/2, 1/4)$. Both sharpness bounds persist; summability replaces pointwise monotonicity. Summing (24) bounds the additional paired block excursion by $C\epsilon u_i^2(\rho+\rho^2)/\beta$. Reduce ρ to fit this and an odd increment inside half the stopping radius. This prevents the block exit and proves the bounds for every future pair.

At each fixed width the Euclidean lengths of the paired increments are summable. The even iterates are Cauchy and the intervening odd increments tend to zero, so both converge to the same point. Continuity and residual contraction give interpolation. The kernel floor, preserved in the weighted neighborhood, gives regularity. These estimates hold at each fixed finite width and extend the finite-horizon approximation to the full trajectory. □

The finite-time transfer and summable tail yield Theorem 6.1.

Proof. First localize to bounded initial a_0, d_0, kernel and drift floors, and the weighted moment event of Lemma F.2. When $|a_0| \geq a_{\min} > 0$, the ODE invariant has radius $R_m = (d_0^2 + 2D_0 a_0^2/\lambda_0^2)^{1/2}$ bounded away from zero. The explicit solution in (22), with initial quantities in place of endpoint quantities, reaches $|r| < \rho/2$ and $d < -3\beta$ after a finite slow time uniformly on this compact set. Finite-time tracking supplies actual entrance to Lemma F.3. Thus entrance is proved, not assumed.

For the i.i.d. law, centering and Chebyshev give a tight residual, $\lambda_0 = \lambda_\infty + O_P(\epsilon)$, $\kappa_0 = O_P(\epsilon)$, and $D_0 = D_\infty + O_P(\epsilon)$. The exceptional event $\lambda_0 = 0$ has probability tending to zero; one may define the step to be zero there to make the algorithm total. Continuous dependence on each fixed time interval replaces $\eta_0^2 D_0$ by $4D_\infty/\lambda_\infty^2$ at cost $O_P(\epsilon)$. Compact support makes the weighted moment event immediate. More generally, a derivative envelope u with $\mathbb{E}u^8 < \infty$ supplies it by the law of large numbers; require also $\mathbb{E}\psi^2 < \infty$. If $\mathrm{Var}(\psi) > 0$, the i.i.d. central limit theorem gives a continuous Gaussian limit for r_0. For any failure probability we can choose $a_{\min} > 0$ small enough to exclude $0 < |a_0| < a_{\min}$ outside that probability. If $\mathrm{Var}(\psi) = 0$, centering implies $\psi = 0$ almost surely: $y \neq 0$ gives a fixed nonzero residual, while $y = 0$ is already an exact stationary interpolant. This covers every case.

The entire trajectory has block displacement $O_P(\epsilon u_i^2)$. The finite derivative-moment family used by $\mathfrak{F}$ and $\mathcal{D}$ has local derivative bounds Cu_i^4 and square-integrable initial values under the eighth-moment envelope. Their endpoint changes are $O_P(\epsilon\overline{u^6})$, and their sampling errors are $O_P(\epsilon)$. The output moment is handled by $\mathbb{E}\psi^2 < \infty$ and its derivative envelope. The endpoint-to-start bounded-Lipschitz distance is $O_P(\epsilon)$, while convergence of the initialized empirical measure to the population law suffices for the remaining term.

At the actual regular scalar GD endpoint the exact identities are

$$mc_1 = \frac{8\kappa^2}{\lambda^3} - \frac{4}{3\lambda^2}\overline{D^3\psi[h,h,h]}, \qquad \mathcal{D} = 4\left(\overline{\|Ah\|^2} - \frac{\kappa^2}{\lambda}\right).$$

Moment convergence and the kernel floor prove the stated joint law there. Polynomial-growth C^4 activations with Gaussian blocks have every required envelope moment; independent centered symmetric readouts provide the two centering identities. The dynamical theorem additionally uses positive drift; for Softplus its nonzero hidden component makes this automatic.

For the final sharpness formula put $I_k := d_k^2 + \eta_0^2 D_0 a_k^2/2$. The leading terms in (25) cancel in ΔI and

$$|\Delta I_k| \leq C\epsilon^2(|r_k| + r_k^2).$$

The finite entrance interval and the geometric tail sum this to $I_\infty = I_0 + O(\epsilon)$. Since $a_\infty = 0$ and $d_\infty < 0$, $d_\infty = -R_m + O(\epsilon)$ on the localized nonstationary event. The kernel and drift concentration estimates replace λ_0, D_0 in this radius by their population values at cost $O_P(\epsilon)$, proving the main theorem's displayed version. The same estimates give the summable vector identity

$$\Delta\theta - v_0\Delta a - \frac{2B_0}{\eta_0 D_0}\Delta d = O\big(\epsilon^2(|r| + r^2)\big),$$

so the actual endpoint also satisfies

$$\theta_\infty = \theta^0 - a_0 v_0 + \frac{2(d_\infty - d_0)}{\eta_0 D_0} B_0 + O_P(\epsilon).$$

All rate statements are understood on the convergence event, whose probability tends to one. No equality of the GD and GF parameter endpoints is asserted. □

F.4 Compatibility of learning rates

Lemma F.4 (Initialization and endpoint calibration agree at theorem precision)**.** *Under the preceding balanced i.i.d. assumptions, let λ_{GF} be the kernel of the analytical GF endpoint from the same initialization. Then $\lambda_{\mathrm{GF}} - \lambda_0 = O_P(m^{-1})$. For fixed d_0, the endpoint- and initialization-calibrated rates differ by $O_P(m^{-1})$ and are covered by the same direct-trajectory theorem at $O_P(m^{-1/2})$ precision.*

Proof. The scalar GF bootstrap gives $\int_0^\infty |r|dt = O_P(1)$ and block motion $O_P(\epsilon u_j)$. Cubic balance and its local derivative bound give $\sup_t |\kappa(t)| = O_P(\epsilon)$. Therefore $\dot{\lambda} = -2rg^\top Hg = -2\epsilon r\kappa$ integrates to $O_P(\epsilon^2)$. The kernel floor proves the rate difference. The deterministic GD proof is uniform for bounded normalized initial detuning; the endpoint-calibrated rate has detuning $d_0 + O_P(\epsilon)$ at initialization. Continuous dependence of the ODE and the proved tail estimates give the conclusions. The separate fixed population rate $2/\lambda_\infty$ is also admissible, with tight random initial detuning $2(\lambda_0 - \lambda_\infty)/(\epsilon\lambda_\infty)$; the limiting system retains this random term. □

F.5 Degenerate tangent drift

Corollary F.5 (Opposite flip signs at the actual GD endpoints)**.** *Both initialization laws of Corollary 5.3 satisfy Theorem 6.1. Their direct GD endpoints have opposite signs of c_1 with probability tending to one, while their population kernels are identical and both tangent-drift coefficients are positive.*

Proof. Both laws are compact, output-centered and third-order balanced, with positive kernel and drift as computed in Corollary 5.3. Apply the endpoint law of Theorem 6.1. Their exact limits are $27(m_4 - 8\log 4)/(512\lambda_\infty^2)$ for $m_4 = 1, 16$, which have opposite nonzero signs because $1 < \log 4 < 2$. □

Proposition F.6 (Zero tangent drift can cause divergent near-critical GD)**.** *The broad GF-selection hypotheses permit a SiLU initialization whose near-critical supercritical GD trajectory diverges.*

Proof. Let $z_c \in (-2, -1)$ be the unique root of $z + 1 + e^z = 0$ and put $b = -1 - z_c = e^{z_c} > 0$. For SiLU, $s(z_c) = -b \neq 0$ and $s'(z_c) = 0$. Use one datum $x = 0$, fixed label y, i.i.d. Gaussian readouts and weights, and bias z_c in every block. This is an i.i.d. subgaussian block law with centered output, polynomial derivative envelopes, and deterministic positive kernel b^2. All hidden gradients are zero exactly, although every parameter is trainable. GF has $r(t) = r_0 e^{-b^2 t}$ and reaches a regular interpolant. In contrast, at $\eta_m = (2 + d/\sqrt{m})/b^2$, $d > 0$,

$$r_{k+1} = -(1 + d/\sqrt{m})r_k.$$

The Gaussian initial residual is nonzero almost surely, hence diverges. Here $Ah = 0$ and $D = c_1 = 0$, identifying positive tangent drift as the nondegeneracy condition that separates this example from Theorem 6.1. □

G Additional experimental details

G.1 Additional coefficient checks

Matrix-free and explicit sample-space evaluations agree to relative c_1 error below 4.3×10^{-14} in 25 cases. A separate 20-case direct loss-derivative check gives maximum scaled c_1 error 1.45×10^{-14} against the residual–Gram calculation.

Of 72 earlier i.i.d. endpoint attempts, 67 interpolate and 65 match the fixed-data population sign. All 19 positive-coefficient endpoints form the projected top-normal branch; among seven negative-coefficient endpoints, six escape and one reaches the step cap.

G.2 Normal–tangent experiments

The two laws of Corollary 5.3 use widths $256, 1024, 4096$, three seeds, target 0.5, initial amplitude 0.2, and detuning $|c_1|(0.4)^2$. All 18 selected endpoints interpolate and match the analytic sign; maximum sampled population-ODE amplitude error falls from 0.000812 to 0.000197 across widths.

For Softplus, tanh, GELU, and SiLU, the same widths and seeds give 36 gradient-flow endpoints and 72 completed endpoint-local trajectories under $0.3/\sqrt{m}$ or $|c_1|(0.4)^2$ detuning. The block law has independent Rademacher readout and uniform hidden bias on $[-1, 1]$; all endpoints reach residual 10^{-8}. Coefficients come from endpoint jets or order-160 population quadrature, independently of the GD paths. Across at most 121 sampled even iterates, maximum endpoint-coefficient amplitude

and normalized sharpness errors under $0.3/\sqrt{m}$ fall from $(0.00442, 0.253)$ to $(0.00102, 0.0664)$ across widths.

Independent checks across 12 coefficient/drift and 24 dense-Hessian cases agree within 10^{-11} for c_1, tangent drift, and actual loss sharpness.

G.3 Trajectories from random initialization

The direct-trajectory grid uses Softplus, target 0.5, widths $256, 1024, 4096$, three seeds, compact Rademacher/uniform blocks or standard Gaussian blocks, and two constant-step policies: initialization calibration with $d_0 = 0.3$ or the population-critical rate. All 36 trajectories proceed directly from initialization and reach $|r| \le 10^{-9}$ within $200\sqrt{m}$ pairs. Their final normalized sharpness is negative in all cases. Thirty-five reach the diagnostic tail criterion $d \le -0.1$, $|r| \le 0.05$; the remaining case also converges. Across those 35, the largest residual contraction ratio is 0.98795.

Mean absolute endpoint coefficient errors fall from 0.011647 to 0.001311 for compact starts and from 0.027960 to 0.001774 for Gaussian starts between widths 256 and 4096. Each initialization is tested under both rate choices; population coefficients use fixed quadrature.

The exact two-sign grid uses the two Softplus laws, three seeds, the same three widths and target 0.5, and fixed initial-kernel detuning $d_0 = 0.3$. All 18 trajectories interpolate to 10^{-9} and match their predicted flip signs. A GF reference is evaluated post hoc to measure the calibration difference.

G.4 Multiple-residual and deep-network experiments

These endpoint-local tests use numerically selected interpolants. Each full and projected pair shares its endpoint, top-normal perturbation, step, and horizon; projection retains the entire fixed normal space $\theta_* + \mathrm{range}(J_*^\top)$.

G.4.1 Multiple-residual ridge networks

The 70 ridge endpoints use $n \in \{16, 64, 256\}$, $d = 16$, $m \in \{512, 2048\}$, Softplus/tanh, and five seeds per cell, plus $n = 64, m = 8192$. Inputs are i.i.d. $N(0, I/d)$; an independent tanh teacher supplies targets. Readouts are Rademacher, hidden weights are uniform on $[-\sqrt{3}, \sqrt{3}]$, and biases on $[-1, 1]$. L-BFGS (at most 400 iterations and a 20-second callback budget) is followed by at most eight backtracked sample-space Newton corrections. All endpoints meet residual RMS 10^{-6}. We use $a_0 = 0.2$, pair-time horizon $T = 3$, and both $\delta = 0.3/\sqrt{m}$ and $\delta = |c_1|(2a_0)^2$.

For the top kernel eigenpair (μ_*, e_*), amplitude is $e_*^\top r/\sqrt{\mu_*}$. The scalar ODE is an empirical prediction here, using the independently computed $D_m = m\,\|P_T \nabla S\|^2$; its error is maximized over at most 49 sampled even iterates. Nine times also use a matrix-free loss-Hessian eigensolve. Projection uses $P_N = J_*^\top (J_* J_*^\top)^{-1} J_*$.

Of 140 comparisons, 130 complete and have smaller final full-GD amplitude. All ten escapes are Softplus at $n = 256$ under coefficient-based detuning, where $\delta \in [1.29, 11.93]$: eight full and two projected trajectories leave radius 100. The main-text table gives the width-scaled detuning grid; the numerical records retain both detunings and all attempts.

G.4.2 Deep endpoint tests

The data are the scikit-learn copy of the UCI Optical Recognition of Handwritten Digits test set (Alpaydin & Kaynak, 1998). We take the first 128 examples of each class, 0 and 1, center and feature-scale the pixels, divide by $\sqrt{64}$, and map labels to $\{-1, 1\}$. The network has three 128-wide hidden layers (41,473 parameters). Across GELU, tanh, and SiLU at seeds $314159, 271828, 161803$, endpoint selection uses 2,000 Adam steps, at most 600 L-BFGS iterations, and at most eight Newton corrections. Three GELU and two SiLU endpoints meet residual 10^{-7}; one SiLU and all three tanh attempts do not.

The five reached endpoints use $\delta = \min\{0.01, \mathrm{gap}(K_*)/(4\mu_*)\}$, $\eta = (2 + \delta)/\lambda_*$, and $p_0 = \min\{0.02, \delta/(\eta\,\|P_T \nabla S\|)\}$ for 1,000 updates from $\theta_* + p_0 v_*$. Final projected/full top-amplitude ra-

Table 2: Every interpolating endpoint in the direction intervention. The final normal/full amplitude ratio measures the damping contrast; the last three columns are the directly computed late-window scores.

Activation	Seed	Normal/full	Necessity	Sufficiency	Random
GELU	314159	33.3	1.00067	1.00051	$-1.2 \cdot 10^{-9}$
GELU	271828	45.5	1.00054	1.00051	$-2.5 \cdot 10^{-9}$
GELU	161803	104.3	1.00102	1.00072	$-2.9 \cdot 10^{-10}$
SiLU	314159	68.6	1.00173	1.00071	$-1.7 \cdot 10^{-9}$
SiLU	161803	573.4	1.00764	1.00021	$-5.2 \cdot 10^{-9}$
GELU	20260925	54.8	1.00199	1.00134	$-1.8 \cdot 10^{-9}$
GELU	20260926	61.7	1.00053	1.00071	$-1.4 \cdot 10^{-9}$
SiLU	20260926	67.9	1.00123	1.00040	$-8.2 \cdot 10^{-9}$

tios are 33.3–573.4. Projected sharpness also falls below threshold; normal motion remains material even though the tangent term accounts for 67.4–93.6% of the full-GD predicted sharpness drop.

The runs use CPU double precision and sample-space or QR projection, without forming a parameter Hessian. Independent dense checks give ridge Hessian-vector relative error below 2.2×10^{-10}, deep sharpness derivative error 1.60×10^{-16}, and deep eigen-residuals below 7.27×10^{-9}.

G.4.3 Intervening on the predicted tangent direction

With the settings of Appendix G.4.2, add seeds $20260925, 20260926$ for GELU/SiLU to the nine original attempts. Eight of 13 endpoints interpolate; three tanh and two SiLU attempts fail, including SiLU seed 20260925 at residual 6.16×10^{-5}. All 40 intervention trajectories from reached endpoints complete 1,000 steps.

The five fixed projections in Section 7 use the full normal projector $P_N = QQ^\top$, $t = P_T \nabla S / \|P_T \nabla S\|$, and an orthogonal random tangent ξ drawn from seed 10^6 plus the initialization seed. Each arm shares its perturbation and step, stops if an iterate leaves radius 0.5 about the endpoint, and is sampled every 50 steps; the final midpoint uses one extra gradient evaluation.

For each method X, define

$$\ell_X := \frac{1}{6} \sum_{k \in \{750,800,850,900,950,1000\}} \log \max\{a_k^X / p_0, 10^{-12}\}, \qquad G := \ell_N - \ell_{\text{full}}.$$

For $G \geq \log 2$, the blocking, restoration, and random-control scores are

$$\frac{\ell_{\text{block}} - \ell_{\text{full}}}{G}, \qquad \frac{\ell_N - \ell_{\text{restore}}}{G}, \qquad \frac{\ell_{\text{random}} - \ell_{\text{full}}}{G}.$$

All eight reached cases have sufficient contrast, scores within 0.2 of $(1, 1, 0)$, and curve discrepancies at most 0.1; blocked final amplitudes are closer to normal-only and restored final amplitudes to full GD.

At step 1,000, normal-plus-t/full amplitude ratios are 0.9945–0.9987, and block-t/normal ratios are 1.0020–1.0493. The maximum sampled curve discrepancy is $0.0323 p_0$. These conditional finite-trajectory comparisons isolate one fixed endpoint direction; sharpness gradients may subsequently rotate.

At an exact eight-example GELU interpolant, dense autodifferentiation and residual derivatives agree on the tangent sharpness gradient within 1.49×10^{-14} relative error; dense and QR projectors agree within 4.76×10^{-14}. Large-run subspace leakage is at most 1.14×10^{-15}. Auxiliary ambient-Hessian power iterations have maximum residual 3.69×10^{-4} and do not enter the amplitude scores.

G.4.4 Residual and numerical-rank sensitivity

At the GELU and SiLU endpoints with seed 314159, baseline QR retains all 256 sample directions. SVD cutoffs through 10^{-6} also retain full rank; the 10^{-4} cutoff retains 253 directions and changes

the normalized tangent direction by 0.098 and 0.043. The smallest relative singular values are 7.02×10^{-5} and 7.30×10^{-5}.

Perturbing these endpoints along $J^{\dagger}\mathbf{1}$ to residual near 10^{-7} changes the full-rank direction by less than 2.0×10^{-8}. Across the two saved, perturbed, and rank-truncated five-arm comparisons, blocking/restoration scores remain within 0.0018 of one and random-control scores within 1.5×10^{-9} of zero. The largest blocking-score change from rank truncation is 3.5×10^{-5}.