NEARLY OPPOSITE GRADIENTS IN SHRINKING REGIONS OF LOOPED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

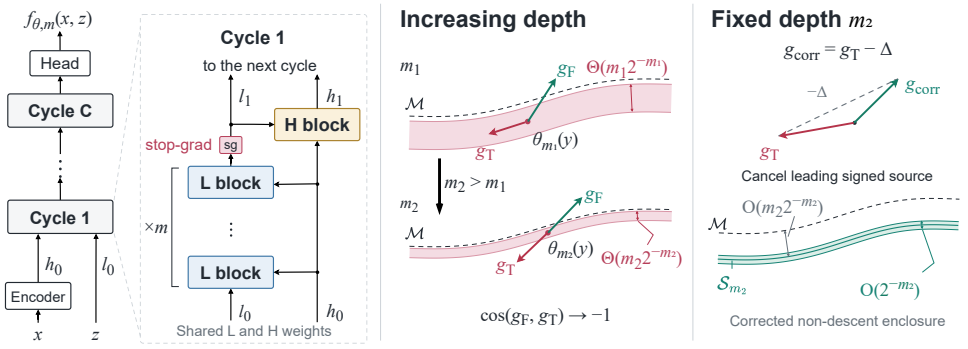
Looped networks increase computational depth by repeatedly applying shared modules, whereas truncated backpropagation differentiates through only a subset of these applications. We study how this mismatch affects gradient direction in a normalized hierarchical loop with m low-level iterations per cycle and truncation of the first low-level group. Increasing depth can make the truncated gradient nearly opposite to the full gradient, even as the reversal region contracts. For a fixed task and neighborhood in the parameter space, the ascent region has thickness $\Theta(m2^{-m})$ and initialization probability $\Theta((m2^{-m})^5)$. A signed expansion shows that omitted parameter contributions accumulate at scale $m2^{-m}$ and compete with the local full gradient. A correction based on data moments and current parameters reduces the remaining thickness to $O(2^{-m})$. We also characterize when discrepancies between cycle contributions persist after low-state convergence. Solver relaxation preserves the limiting alignment, whereas suitable block weights can eliminate an ascent interval. We verify finite-depth boundary predictions in a nonlinear residual family. At trained looped-Transformer checkpoints, restoring early contributions achieves higher cosine similarity with the full gradient than any nonnegative block rescaling and produces larger cross-entropy reductions under matched updates. Thus, greater loop depth can reduce the prevalence of gradient reversal without reducing its severity.

1 INTRODUCTION

Looped networks increase computational depth by reusing modules with shared parameters across successive steps. This design enables iterative computation in Universal Transformers (Dehghani et al., 2019), recurrent-depth language models (Geiping et al., 2025; Saunshi et al., 2025), and hierarchical reasoning architectures (Wang et al., 2025; 2026). Every execution adds a new path for shared parameters to influence the final loss. To compute parameter updates efficiently, truncated backpropagation keeps only a subset of these paths (Williams & Peng, 1990). For instance, HRM-Text executes eight module updates across two cycles, whereas tracks gradients through only the final five updates (Wang et al., 2026). Thus, the forward pass executes a deeper computation than the backward pass differentiates through during optimization (Figure 1(a)). In this work, we study how this depth mismatch alters the direction of parameter updates.

A more accurate forward state does not guarantee that a truncated update will reduce the loss. The norm of the full gradient is small near stationary regions. The sign of the inner product between the truncated and full gradient can be reversed there, even if their individual contributions decay with low-level update depth m. This behavior raises a key concern: how sharply can the two gradients diverge, and where in parameter space does this conflict cause unintended ascent? The limit $m \to \infty$ is taken at fixed high-level cycles C to separate equilibrium behavior from finite-depth transients (Section 2). The forward error is the difference between the finite-depth output and its infinite-depth limit (Section 3). For the normalized hierarchical loop in our geometric analysis(Section 2.3), we identify those points in the parameter space where gradient cosine similarity approaches -1 as depth grows, even while the forward error remains bounded at $O(2^{-m})$. The model has two trainable 2×2 matrices, so its parameter space is $\mathbb{R}^{2\times2} \times \mathbb{R}^{2\times2} \cong \mathbb{R}^8$. Within a fixed neighborhood of this eight-dimensional parameter space, the ascent region maintains a thickness of $\Theta(m2^{-m})$ (Section 2.3) and an initialization probability of $\Theta((m2^{-m})^5)$ as depth m increases (Theorem 3.2, Figure 1(b)). Therefore, greater depth makes ascent directions increasingly rare, even as the worst-case gradient

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.



(a) Shared-loop architecture (b) Severity vs. extent (c) Structural correction

Figure 1: Looped computation and gradient truncation. (a) Truncated backpropagation stops derivatives through the first low-level group. (b) As loop depth increases, the truncated gradient can remain nearly opposite to the full gradient, although this reversal is confined to a narrower region of parameter space. (c) Applying Δ, a correction for the leading omitted terms, reduces the non-descent region.

reversal becomes more severe. Our theoretical analysis identifies the exact gradient component that causes both phenomena.

Inside the omitted low-level loop, state errors at each step multiply with sensitivity terms carried across the remaining computation. This interaction generates an error term of order $m2^{-m}$ that directly opposes the full gradient near the critical set. We derive the direction of this error from the loop structure and data distribution, pinpointing where the gradient reverses. This expansion also provides a practical remedy: a targeted correction based on data moments (Assumption 2.1) and current parameters eliminates the primary discrepancy (Section 3; Appendix A.7), shrinking the error bound to $O(2^{-m})$ (Figure 1(c)). We extend this analysis to general hierarchical loops by comparing how retained and omitted cycles contribute to parameter updates. These contributions diverge even when low-level loops fully converge (Section 2; Appendix B.3), since each cycle operates at a distinct high-level state and influences a different phase of downstream execution. This contrast highlights a fundamental difference between modifying the forward solver and modifying the gradient update: solver relaxation preserves the limiting full/truncated gradients (Section 2; Appendix B.4), while block reweighting ensures descent throughout the task family (Section 4; Appendix B.5).

We evaluate this explanation in both a solvable model and trained looped Transformers. In the solvable model, we verify gradient reversal and predict its ascent boundaries. On trained Transformer checkpoints, derivative restoration matches the full gradient more closely and yields larger loss reductions than weight rescaling (Section 5). These results show that forward refinement and backward correction guide model updates in fundamentally different ways.

Our contributions are threefold: ♠ **Sharp geometry of reversal.** We derive a signed gradient expansion and prove that the ascent region has thickness $\Theta(m2^{-m})$ and initialization probability $\Theta((m2^{-m})^5)$, while identifying depth-dependent centers where $\cos(g_{\mathrm{F},m}, g_{\mathrm{T},m}) \to -1$. We also construct a structural correction that contracts the thickness to $O(2^{-m})$. ♥ **General hierarchical loops.** We derive a cycle-wise decomposition of the limiting alignment and prove that solver relaxation preserves it, while block weights above an exact threshold ensure descent across an affine task family. ♣ **Evidence.** We validate the finite-depth predictions in a nonlinear loop and show that restoring omitted derivatives improves alignment and loss reduction more than block rescaling or additional forward iterations at trained looped-Transformer checkpoints.

2 Preliminaries

2.1 Forward computation and objective

The loop maintains high and low states h and l over $C \geq 2$ high-level cycles, with $m \geq 1$ low-level updates per cycle (Wang et al., 2025; 2026). The encoder, high-level module, low-level module, and head have disjoint parameter blocks $\theta = (\theta_E, \theta_H, \theta_L, \theta_{\mathrm{head}})$. Recurrent parameters are shared

across applications. Given input x, target y, and an initial low state z independent of θ, the forward computation for $c = 1, \dots, C$ is

$$l_c = \mathcal{L}_{\theta_L}^m(l_{c-1}; h_{c-1}), \quad h_c = \mathcal{H}_{\theta_H}(h_{c-1}, l_c), \quad f_{\theta,m}(x,z) = \mathrm{Head}_{\theta_{\mathrm{head}}}(h_C), \tag{1}$$

where $h_0 = \mathrm{Enc}_{\theta_E}(x)$ and $l_0 = z$, as illustrated in Figure 1(a). For each high-level cycle, the low-level group $\mathcal{L}_\theta^m$ is the block of m updates performed while the high state in that cycle is held fixed. For a fixed data distribution over (x, z, y) and a sample loss ℓ, the objective is $F_m(\theta) = \mathbb{E}\big[\ell\big(f_{\theta,m}(x,z), y\big)\big]$. We assume that the sample loss is continuously differentiable with repsect to the parameters and the gradient is uniformly bounded. The uniform bound is moreover assumed to be integrable under the data distribution (Appendix B.3).

2.2 Gradients and alignment

For full backpropagation, we compute the full gradient $g_{\mathrm{F},m} = \nabla F_m(\theta)$ (Baydin et al., 2018). For analytical tractability, we consider a specific form of truncated backpropagation (Williams & Peng, 1990) in which gradients are truncated at the output l_1 of the first low-level group. Specifically, we obtain the truncated gradient $g_{\mathrm{T},m}$ by applying $l_1 \leftarrow \mathrm{sg}(l_1)$ at every subsequent use of this output, where sg is the stop-gradient operator. We call this backward rule the mask. Both gradients use the same parameters, samples, and forward states. The Frobenius product is adopted for the inner products of mareices. We denote derivatives by D and use induced operator norms for derivative maps. For $\eta > 0$, we define

$$\theta^+ = \theta - \eta g_{\mathrm{T},m}, \qquad D_m(\theta) = \langle g_{\mathrm{F},m}, g_{\mathrm{T},m} \rangle. \tag{2}$$

We call $\{\theta : D_m(\theta) < 0\}$ the ascent region. The low-parameter block is the θ_L coordinate block of a full parameter. The group contribution c_i is the gradient of F_m with respect to a temporary low-parameter copy $\theta_{L,i}$ assigned to the i-th low-level group. We keep $\theta_{L,i}$ shared across that group's m updates, treating the copies for different groups as independent while differentiating, and then evaluate at $\theta_{L,1} = \cdots = \theta_{L,C} = \theta_L$ (Pascanu et al., 2013). Under this block order, we define a matrix $P := \mathrm{diag}(I_E, I_H, r_C I_L, I_{\mathrm{head}})$, with each identity acting on its corresponding block. The matrix P is the reference operator for the equal-contribution case: if the C groups contribute equally, omitting the first group changes only the low-parameter gradient by this factor. We also consider the situation when $m \to \infty$ at fixed C, data, and mask.

2.3 Normalized model and local geometry

We consider two-dimensional states and shared trainable matrices $\theta = (A, B) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2} \cong \mathbb{R}^8$ for simplicity. Let $e(x)$ be the direction vector $(\cos x, \sin x)^\top$ and denote $N(v) = v/\|v\|$ as the normolization on nonzero vectors. Following (1), for a scalar input X and $\kappa \in (0,1)$, the model is defined as

$$h_0 = e(X), \quad l_0 = e(0), \quad l \leftarrow N(B(l+h)), \quad h \leftarrow N(A(h+l)). \tag{3}$$

The corresponding loss is then $F_m = \mathbb{E}[1 - \langle h_C, e(\kappa X)\rangle]$. We now state the data-moment condition used in the local analysis.

Assumption 2.1 (Data moments)**.** For the fixed $\kappa \in (0,1)$, the input X has a distribution symmetric about zero, meaning that X and $-X$ have the same distribution supported on $[-a_0, a_0]$ with $a_0 < \pi$. Set $r = \sin((1-\kappa)X)$ and $c = \cos((1-\kappa)X)$. The moments satisfy

$$\begin{aligned} &\mathbb{E}[r \sin 2X] = 0, \quad B_0 = \mathbb{E}[2Xr\cos 2X] \neq 0, \\ &J = \mathbb{E}[r \sin 4X] > 0, \quad K = \mathbb{E}[c \sin^2 2X], \quad K + J > 0, \\ &M_2 = \begin{pmatrix} \mathbb{E}c & \mathbb{E}[c\cos 2X] \\ \mathbb{E}[c\cos 2X] & \mathbb{E}[c\cos^2 2X] - J \end{pmatrix} \succ 0. \end{aligned} \tag{4}$$

With this fixed data distribution, we next describe the parameter set on which the local geometry is analyzed.

Let R_φ denote a planar rotation through angle φ. For sufficiently small $\varepsilon > 0$, define the set

$$\mathcal{M} = \{(\alpha \mathsf{R}_\varphi, \beta \mathsf{R}_{-\varphi}) : \alpha, \beta > 0,\ |\varphi| < \varepsilon\}. \tag{5}$$

To keep the local geometry manageable, we focus on an interior compact portion $\mathcal{M}_0$ of this surface and a slightly larger portion $\mathcal{M}_1$, restricting attention to a small neighborhood U of $\mathcal{M}_0$. In this neighborhood, each parameter point is described by a nearby point $\bar{\theta}$ on the surface and a displacement n perpendicular to it. We use $H(\bar{\theta}) = \nabla^2 F_\infty(\bar{\theta})$ for the Hessian of the limiting loss and $H_N(\bar{\theta})$ for its restriction to perpendicular directions, hence thickness is measured by $\|n\|$. The formal definitons with precise projection and normal-coordinate conditions are stated in Appendix A.

2.4 Equilibrium limits and interventions

For the equilibrium analysis, we assume a unique conditional low equilibrium $e_\theta(h)$ and local uniform convergence of the low iterates and their first derivatives to this map and its derivatives, respectively. The corresponding high-level transition is $T(h) = \mathcal{H}_\theta(h, e_\theta(h))$. The remaining maps are continuously differentiable near the visited and limiting states, with integrable bounds on their derivatives (Appendix B.3). The limiting alignment is defined to be $D_\infty = \langle g_{\mathrm{F},\infty}, g_{\mathrm{T},\infty} \rangle$. We define the low-block discrepancy as $e_{\mathrm{L}} = g_{\mathrm{T},\infty,\mathrm{L}} - r_C g_{\mathrm{F},\infty,\mathrm{L}}$ and the omitted encoder contribution $b_{\mathrm{E}} = g_{\mathrm{F},\infty,\mathrm{E}} - g_{\mathrm{T},\infty,\mathrm{E}}$, with $b_{\mathrm{E}} = 0$ for a fixed encoder. We consider task families with squared loss and affine targets $y(a) = y_0 + a y_1$, $a \in \mathbb{R}$. An ascent interval is an interval of a with negative gradient alignment.

3 Nearly Opposite Gradients in a Shrinking Region

We begin with a signed gradient expansion that identifies the contribution omitted by truncation. The main theorem then uses this expansion to locate parameter points with nearly opposite gradients and determine the size of the ascent region. A correction tests the proposed mechanism by canceling the identified contribution.

3.1 The omitted contribution near a critical set

On the set $\mathcal{M}$, normalization removes the scale factors, and the rotations cancel in the equilibrium high-level update. Under Assumption 2.1, $\mathcal{M}$ is a critical set of the limiting loss. If the low-level groups contributed equally, we will have $g_{\mathrm{T},m} = P g_{\mathrm{F},m}$ by truncation. This positive block rescaling preserves descent. The next proposition quantifies the signed discrepancy from this reference. By setting $p = 2^{-m}$, $s = mp$, and $f(\bar{\theta}) = \lim_{m\to\infty} g_{\mathrm{F},m}(\bar{\theta})/s$, we have

Proposition 3.1 (Signed gradient expansion)**.** *Under Assumption 2.1, $\mathcal{M}$ is a critical set of F_∞, and $H_N(y)$ is invertible on the chosen patches. The vector $f(y)$ is nonzero, normal to $\mathcal{M}$, and supported on the low-parameter block. For any fixed K_0, the expansions*

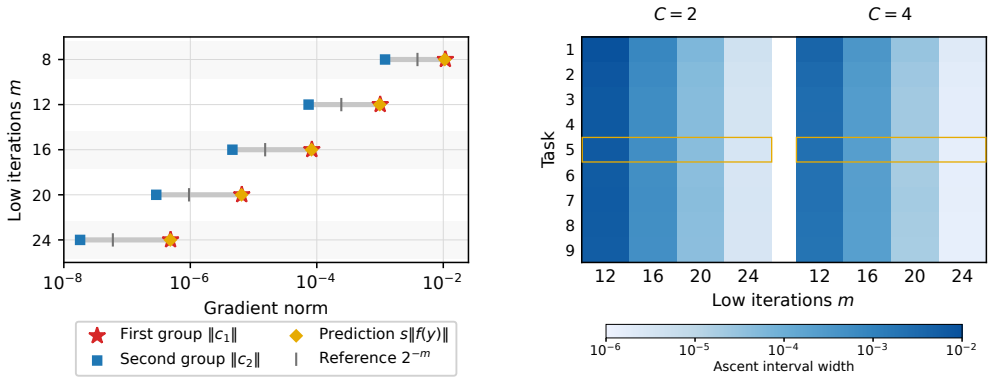
$$g_{\mathrm{F},m}(y+n) = H(y)n + sf(y) + O(p), \qquad g_{\mathrm{T},m}(y+n) = PH(y)n + O(p) \tag{6}$$

hold uniformly for $y \in \mathcal{M}_1$ and normal displacements $\|n\| \le K_0 s$.

The term $H(\bar{\theta})n$ gives the first-order change in the limiting gradient under a normal displacement n. Truncation removes the leading contribution $sf(\bar{\theta})$ from the first low-level group, yielding $g_{\mathrm{T},m} - P g_{\mathrm{F},m} = -sPf(\bar{\theta}) + O(p)$. At normal distances of order s, both leading terms have magnitude of order s, hence the omission of $sf(\bar{\theta})$ can change the update direction. We derive the Hessian and coefficient from the data moments and current matrices in Appendices A.4 and A.2.

The factor m arises from repeated use of the shared parameters. At $\bar{\theta} \in \mathcal{M}$, the leading state mismatch at visit j of the omitted group has order 2^{-j}. Propagation through the remaining visits contributes a factor of order $2^{-(m-j)}$. Each product therefore has scale p, and the data moments give these m parameter contributions a common nonzero leading coefficient, yielding $sf(\bar{\theta})$. Subsequent low-level groups start within $O(p)$ of their conditional equilibria. After averaging over the data, their contributions and the direct high-parameter terms are $O(p)$, the mask hence removes the entire leading term of order s (Appendix A.5). The factor m also appears in sensitivity-error bounds for contractive iteration (Grazzi et al., 2020, Proposition 2.1).

We measure gradient contributions across nine tasks, two cycle counts, and three fixed parameter points on $\mathcal{M}$. The first-group norm follows the prediction $s\|f(\bar{\theta})\|$, while the second has scale p (Figure 2(a)). We specified these predictions before evaluation. The protocol and independent numerical checks are provided in Appendix D.2.



(a) Omitted and retained contributions

(b) Measured widths across nine tasks

Figure 2: Measured gradient contributions and ascent widths. (a) Paired norms of the first and second low-group contributions. (b) Measured ascent widths along the specified normal lines for all nine tasks and both cycle counts, shown on a common logarithmic color scale. The outlined row corresponds to the uniform task in panel (a).

3.2 Stronger reversal in a smaller region

Proposition 3.1 describes the competition between the local full gradient and the omitted contribution. The following theorem locates parameter points where this competition makes the gradients nearly opposite, giving matching bounds on the size of the ascent region within a fixed neighborhood.

Theorem 3.2 (Sharp geometry of ascent directions)**.** *Under Assumption 2.1, there exist a fixed neighborhood U of the chosen patch and constants $K, c_{\mathrm{in}}, c_{\mathrm{asc}} > 0$, all independent of m, such that the following properties hold for all sufficiently large m. Every $\theta \in U$ with $\operatorname{dist}(\theta, \mathcal{M}) \geq Ks$ satisfies $D_m(\theta) > 0$. For each $\bar{\theta} \in \mathcal{M}_0$, consider the normal ball of radius $c_{\mathrm{in}} s$ centered at*

$$\theta_m\left(\bar{\theta}\right) = \bar{\theta} - \frac{s}{2} H_N\left(\bar{\theta}\right)^{-1} f\left(\bar{\theta}\right). \tag{7}$$

Every point in this ball satisfies $D_m(\theta) \leq -c_{\mathrm{asc}} s^2$. At the center, $\cos(g_{\mathrm{F},m}, g_{\mathrm{T},m}) \to -1$, while the forward output differs from its infinite-depth value by $O(p)$.

For any fixed initialization density supported on U, bounded above there, and bounded below by a positive constant near $\mathcal{M}_0$,

$$\Pr\{\theta_0 \in U : D_m(\theta_0) < 0\} = \Theta(s^5) = \Theta\big((m2^{-m})^5\big). \tag{8}$$

The theorem distinguishes parameter points held fixed from centers that depend on depth. Every fixed parameter point in U at a positive distance from $\mathcal{M}$ eventually has positive alignment. The centers $\theta_m(\bar{\theta})$ approach $\mathcal{M}$ at scale s, with both gradient norms tending to zero and their cosine tending to -1. At these centers, the omitted contribution remains comparable to the full gradient. For the same fixed task and neighborhood, the ascent region has thickness $\Theta(s)$ and initialization probability $\Theta(s^5)$ under the stated density conditions. The exponent five corresponds to the five normal directions in parameter space. More accurate forward states can therefore coexist with stronger reversal in a smaller region. We show the complete proof in Appendix A.6.

We also measure the ascent interval along the normal line through $\bar{\theta}$ in the direction $-H_N(\bar{\theta})^{-1} f(\bar{\theta})$. The signed expansion predicts a leading-order width of $s\|H_N(\bar{\theta})^{-1} f(\bar{\theta})\|$ along this line. Figure 2(b) reports the measured widths for all nine tasks and both cycle counts, computed from $g_{\mathrm{F},m}$ and $g_{\mathrm{T},m}$ with respect to all eight parameters. The width decreases with depth in all combinations of task and cycle count. Particularly, for the uniform task with $C = 2$, it decreases from 6.28×10^{-3} at $m = 12$ to 3.25×10^{-6} at $m = 24$.

3.3 A correction that tests the explanation

The signed expansion identifies how to remove the leading cause of reversal. We form the corrected gradient $g_{\mathrm{corr},m}$ by adding $sPf(\bar{\theta})$ after projecting out its component parallel to the current low

Table 1: Ablation of the signed correction at the predicted centers.

	$m = 12$		$m = 16$	
Update	Cosine ↑	ΔF (10^{-10}) ↓	Cosine ↑	ΔF (10^{-12}) ↓
Truncated	-0.328 ± 0.263	$+0.49 \pm 0.42$	-0.592 ± 0.188	$+0.56 \pm 0.25$
Low-block rescaling	-0.498 ± 0.308	$+0.74 \pm 0.48$	-0.719 ± 0.187	$+0.67 \pm 0.26$
Reversed correction	-0.826 ± 0.080	$+1.25 \pm 0.41$	-0.896 ± 0.051	$+0.84 \pm 0.27$
Structural correction	0.988 ± 0.008	-1.50 ± 0.48	0.992 ± 0.005	-0.93 ± 0.29
Full gradient	1.000 ± 0.000	-1.51 ± 0.48	1.000 ± 0.000	-0.94 ± 0.29

matrix. This projection respects the scale invariance of the normalized low-level module. We compute the correction from the data moments and current parameters. The construction in Appendix A.7 yields $g_{\mathrm{corr},m} = P g_{\mathrm{F},m} + O(p)$ for $\|n\| = O(s)$, reducing the discrepancy from order s to order p.

The following theorem bounds the ascent region that remains after this cancellation.

Theorem 3.3 (A smaller ascent region after correction)**.** *In the setting of Theorem 3.2, the following implication holds in a possibly smaller fixed neighborhood U for all sufficiently large m:*

$$\langle g_{\mathrm{F},m}(\theta), g_{\mathrm{corr},m}(\theta)\rangle \le 0 \implies \mathrm{dist}(\theta, \mathcal{S}_m) \le K' p, \tag{9}$$

where $\mathcal{S}_m = \{\bar\theta - sH_N\left(\bar\theta\right)^{-1} f\left(\bar\theta\right) : \bar\theta \in \mathcal{M}_1\}$. Under the bounded-density assumption of Theorem 3.2, we also have $\Pr\{\theta_0 \in U : \langle g_{\mathrm{F},m}(\theta_0), g_{\mathrm{corr},m}(\theta_0)\rangle \le 0\} = O(p^5)$. Consequently, the factor m^5 is removed from the probability bound.

The set $\mathcal{S}_m$ describes the zeros of the leading full-gradient approximation. The correction confines nonpositive alignment to a region of thickness $O(p)$ around this shifted set, removing the factor m from the thickness bound. This reduction tests whether the identified signed contribution accounts for the original scale $m2^{-m}$. We provide the proof and explicit construction in Appendix A.7. We evaluate the correction and its effect on stationary-point displacement in Section 5.1 with corresponding formulas given in Appendix A.8.

We evaluate the correction at the predicted centers in (7), using the same nine tasks, two cycle counts, and three manifold locations as in the contribution study. For each update vector $d \neq 0$, we use $\theta^+ = \theta - \eta d/\|d\|$ and report the recomputed loss change $\Delta F = F_m(\theta^+) - F_m(\theta)$. Table 1 compares the correction with two controls. Low-block rescaling multiplies the retained low gradient by $C/(C-1)$, whereas reversed correction applies the proposed adjustment with the opposite sign. The structural correction yields positive alignment and reduces the loss at all parameter points across the two depths. At each depth, both controls have negative mean cosine similarity and positive mean loss change. This comparison directly tests the predicted sign of the omitted contribution.

4 Which Gradient Contributions Survive More Looping?

In the normalized model, an omitted contribution that decays as $m2^{-m}$ explains gradient reversal near the critical set. The underlying cancellations depend on the recurrence and data moments. We now consider general recurrent modules under the same shared high-low schedule and first-group mask from Section 2. Cycle contributions reveal which discrepancies persist as low-level depth increases and which interventions can change the limiting update.

4.1 Why low-state convergence can leave a gradient discrepancy

The low-parameter contribution of a cycle includes its local parameter derivative and all subsequent paths to the loss. We show in Lemma B.1 the finite-depth discrepancy from positive low-block rescaling as

$$g_{\mathrm{T},m,\mathrm{L}} - r_C g_{\mathrm{F},m,\mathrm{L}} = \frac{1}{C}\sum_{i=2}^{C}(c_i - c_1). \tag{10}$$

Even after low-state convergence, cycles can receive different high states, and earlier parameter responses propagate through more high-level transitions. The next proposition gives the limiting alignment and a sufficient condition for the low-block discrepancy to vanish.

Proposition 4.1 (Cyclewise gradient discrepancy at equilibrium)**.** *Under the equilibrium assumptions in Section 2, let $h_i = T(h_{i-1})$ be the limiting high-state trajectory. The low-block discrepancy satisfies*

$$e_{\mathrm{L}} = \frac{1}{C}\sum_{i=1}^{C-1}(C-i)(c_{i+1,\infty} - c_{i,\infty}), \tag{11}$$

and the full-parameter alignment is

$$D_\infty = \langle g_{\mathrm{F},\infty}, P g_{\mathrm{F},\infty}\rangle + \langle g_{\mathrm{F},\infty,\mathrm{L}}, e_{\mathrm{L}}\rangle - \langle g_{\mathrm{F},\infty,\mathrm{E}}, b_{\mathrm{E}}\rangle. \tag{12}$$

If, for almost every sample,

$$D_{\theta_{\mathrm{L}}} T(h_i) = D_h T(h_i) D_{\theta_{\mathrm{L}}} T(h_{i-1}), \qquad i = 1, \ldots, C-1, \tag{13}$$

then $e_{\mathrm{L}} = 0$. For the partial parameter derivative of T, we hold the high-state argument fixed. If $b_{\mathrm{E}} = 0$ also holds, then $D_\infty > 0$ whenever $g_{\mathrm{F},\infty} \neq 0$.

Condition (13) compares the parameter response introduced in a cycle with the response from the previous cycle after one further high-level transition. Agreement between these responses eliminates the low-block discrepancy, while the encoder has a separate omitted path through the equilibrium of the first low-level group. For a nonzero limiting full gradient, (12) expresses the alignment as a positive quadratic term plus two signed discrepancy terms. The limiting update direction gives first-order ascent when the discrepancy terms have a negative sum whose magnitude exceeds the positive term. This decomposition extends the comparison in Section 3 to loops whose discrepancy can persist after low-state convergence. Appendix B.3 provides the response formulas and proof, including cancellations that yield $e_{\mathrm{L}} = 0$ even when individual responses differ.

4.2 Which interventions change the update direction?

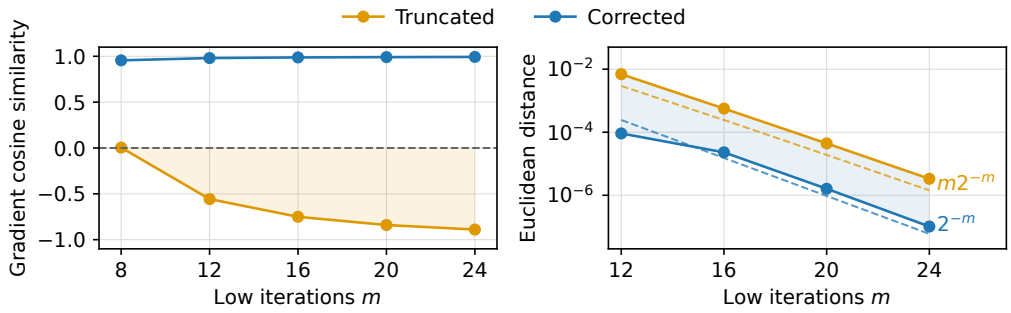
Solver relaxation replaces each low-level update with $(1-\omega)l + \omega\mathcal{L}_\theta(l; h)$ for $\omega \in (0, 1]$. This changes the rate at which a low-level group approaches equilibrium (Mann, 1953). If the low-level map is contractive on a common convex invariant domain and the derivative-convergence conditions hold, every fixed $\omega \in (0, 1]$ preserves the equilibrium map and its derivatives (Grazzi et al., 2020). Proposition B.2 in Appendix B.4 shows that the limiting full and truncated gradients, and hence e_{L}, b_{E}, and D_∞, remain unchanged. Relaxation can shift the depth at which an ascent interval appears. For each task with strictly negative limiting alignment, the truncated update is an ascent direction at all sufficiently large depths. Corollary B.3 provides a finite-depth error bound for this prediction.

Block reweighting assigns weight $\rho \geq 0$ to the exactly retained high and head blocks and one to all other blocks. The threshold ρ_* is the infimum of weights that give nonnegative alignment for every a, or ∞ if none exists. The joint high and head gradient is affine in a; it has full-rank coverage when its constant and linear coefficient vectors are linearly independent. It changes the balance in (12). In the affine task family defined in Section 2, the exactly retained high and head blocks contribute a nonnegative quadratic function of a to the alignment. Under the full-rank coverage condition, a sufficiently large weight on these blocks guarantees descent for every task in the family. Proposition B.4 in Appendix B.5 gives the exact threshold ρ_*. Every weight $\rho > \rho_*$ yields strict descent throughout the family, whereas every $0 \leq \rho < \rho_*$ leaves at least one finite task coordinate with negative alignment. This threshold determines whether reweighting closes the ascent interval at fixed parameters and forward depth.

Derivative restoration reinstates selected backward paths removed by the mask and leaves the forward computation unchanged. It changes the directions available to the update. With two cycles, the full low-block gradient is $c_1 + c_2$, whereas the truncated gradient is c_2. Scalar reweighting keeps this block in $\mathrm{span}(c_2)$. If c_1 has a component outside this span, restoration recovers a direction that scalar reweighting cannot produce. This finite-depth identity motivates the comparison at trained-model checkpoints in Section 5.3. The nonlinear experiment in Section 5.2 precedes that comparison and tests how solver relaxation shifts ascent boundaries and how retained-block weights close the corresponding intervals.

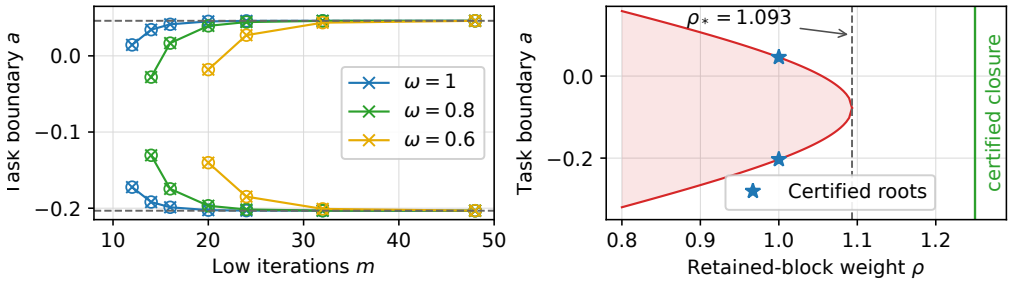
5 Experiments

We evaluate the theoretical predictions in three settings: (i) gradient reversal and its correction in the normalized model (Proposition 3.1, Theorem 3.2), (ii) finite-depth ascent intervals in a nonlinear



(a) Predicted center (b) Stationary-point distance

Figure 3: Gradient reversal intensifies as its geometric scale contracts. Results are shown for the uniform task ($\kappa = 1/2, C = 2$). (a) Gradient cosine similarity at the center specified by Theorem 3.2. (b) Euclidean distance from the full-gradient stationary point on the symmetric diagonal submanifold.



(a) Relaxation shifts boundaries (b) Reweighting closes the interval

Figure 4: Distinct effects of solver relaxation and block reweighting. (a) Ascent boundaries across three relaxation parameters (dashed: shared equilibrium; circles: predictions; crosses: differentiation). (b) Boundaries vs. retained-block weight ($m = 24, \omega = 1$). Exact arithmetic certifies $\rho = 1$ roots and family-wide positive alignment at $\rho = 1.25$.

residual cell, and (iii) gradient interventions at trained looped-Transformer checkpoints. Appendix D details the experimental protocols and provides supplementary results.

5.1 Gradient Reversal and Structural Correction

Every fixed point at distance $\|n\| > 0$ from $\mathcal{M}$ satisfies $D_m > 0$ for large m across enough configurations of the normalized model (3), confirming that fixed configurations eventually leave the ascent region. In contrast, along theoretical centers $\theta_m(y)$ (7), gradient cosine similarity drops from 0.004 at $m = 8$ to -0.889 at $m = 24$, despite forward error decaying at rate $O(2^{-m})$ (Figure 3(a)). On the symmetric diagonal submanifold, truncated updates shift stationary points at the predicted scale $m2^{-m}$, whereas our structural correction reduces this shift to $O(2^{-m})$ (Figure 3(b)).

5.2 Boundary Prediction in Nonlinear Dynamics

We evaluate a 74-parameter Leaky-ReLU recurrent cell across more than 900 affine task families $y(a) = y_0 + ay_1$. Our finite-depth predictions match numerical differentiation to within machine precision (10^{-15}), whereas equilibrium approximations misclassify 40 intermediate ascent intervals (Figure 4(a)). The results highlight a fundamental difference between solver relaxation and block reweighting. Solver relaxation delays the onset of ascent from $m = 12$ to $m = 20$, but all trajectories still converge to the same equilibrium interval (Proposition B.2). In contrast, block reweighting removes the ascent region entirely. The weights $\rho > \rho_*$ on retained blocks ensure descent across all task directions (Figure 4(b)). Empirical tests over around 120 updates confirm this behavior, where $\rho = 1.2\rho_*$ consistently yields descent, while $\rho = 0.8\rho_*$ causes ascent.

Table 2: Restoring omitted contributions improves updates of the trained Transformer. Each row identifies a checkpoint. Left: held-out gradient cosine similarity at $m = 6$ (↑). Right: recomputed cross-entropy change ΔF ($\times 10^{-5}$, ↓) after a step with norm 10^{-4}. "Rest. $\mathcal{L}$": restoration of the early L-group sensitivities.

Seed/Step	**Gradient Cosine Similarity** (↑)				**Cross-Entropy Change** ΔF (↓)		
	Truncated	$\mathcal{L} \times 2$	Opt. Scale	Rest. $\mathcal{L}$	Truncated	Depth 12	Rest. $\mathcal{L}$
9011/200	0.359	0.401	0.490	0.990	−4.72	+0.52	−13.00
9011/600	0.287	0.305	0.333	0.985	−5.11	−1.25	−17.54
9012/200	0.594	0.641	0.743	0.953	−3.68	−1.17	−5.89
9012/600	0.448	0.475	0.531	0.964	−5.12	−0.77	−11.02
9013/200	0.314	0.358	0.454	0.995	−7.77	−0.92	−24.57
9013/600	0.831	0.823	0.837	0.994	−9.55	−3.00	−11.42

5.3 Subspace deficit in looped Transformers

Table 2 evaluates gradient interventions on a looped Transformer trained for next-token prediction on Tiny Shakespeare. Since early sensitivities fall outside the span of later ones ($c_1 \notin \mathrm{span}(c_2)$), scalar reweighting cannot recover the missing update directions. Oracle nonnegative rescaling hence achieves only moderate cosine similarity (0.33–0.84) with the full gradient, whereas derivative restoration (Rest. $\mathcal{L}$) reaches 0.95–0.99.

Under matched step sizes $\eta = 10^{-4}$, a deeper forward pass $m = 12$ fails to fix this alignment deficit and can even harm training progress, as seen at checkpoint 9011/200. In contrast, derivative restoration delivers the largest cross-entropy reduction across all checkpoints, achieving up to $3.5\times$ the descent rate of standard truncation. These findings demonstrate that recovering backward paths benefits optimization far more than extending forward depth.

6 Related Work

Looped architectures and truncated backpropagation. Looped Transformers (Dehghani et al., 2019; Giannou et al., 2023) and hierarchical reasoning models (Wang et al., 2025; 2026; Geiping et al., 2025) increase depth by reusing parameter blocks, often using truncated backpropagation to save memory (Williams & Peng, 1990; Pascanu et al., 2013). Prior work analyzes truncation via norm bounds, compensation, or state stability (Tallec & Ollivier, 2017; Aicher et al., 2020; Miller & Hardt, 2019). In contrast, we show that omitted early sensitivities near critical manifolds cause severe gradient reversal ($\cos \to -1$), even as ascent volumes contract at rate $\Theta((m2^{-m})^5)$.

Equilibrium differentiation and backward interventions. For deep equilibrium models and truncated hypergradients, forward convergence does not guarantee accurate backward derivatives (Bai et al., 2019; Grazzi et al., 2020; Bolte et al., 2023; Geng et al., 2021). Similarly, hierarchical loops retain cycle discrepancies when operation at distinct high states yields non-overlapping sensitivity spans ($c_1 \notin \mathrm{span}(c_2)$). Thus, forward relaxation preserves limiting ascent boundaries, while backward interventions (e.g., block reweighting, derivative restoration) mitigate the directional deficit. Due to space constraints, an extended discussion of related work is provided in Appendix C.

7 Conclusion

Greater loop depth can strengthen gradient reversal while reducing the region in which it occurs. In the normalized hierarchical model, a signed sum of omitted parameter contributions explains both the nearly opposite gradients and the region thickness $\Theta(m2^{-m})$. The same analysis yields a correction that reduces the remaining thickness to $O(2^{-m})$. For more general hierarchical loops, comparing retained and omitted cycle contributions explains why low-state convergence can leave a directional discrepancy and why modifications of the solver, block weights, and backward paths have different effects. Finite-depth predictions in nonlinear loops and matched updates at trained Transformer checkpoints support this interpretation. Overall, the effect of additional loops depends not only on the accuracy of the forward states but also on which parameter contributions the backward pass retains.

Reproducibility Statement

The manuscript source package includes complete proofs, figure data, numerical checks of the claims, and build instructions. The accompanying experiment workspace contains fixed protocols, predictions, checkpoints, and CPU reproduction scripts. Appendix D distinguishes theorem checks, finite-depth predictions, interventions, and numerical certification. We verified the figures and reported counts against the experiment records. All figures are generated from coordinates embedded directly in the LaTeX source.

AI Use Statement

OpenAI Codex was used for literature discovery and comparison of primary sources, development and verification of mathematical derivations, implementation of CPU experiments and numerical audits, analysis of results, and drafting and editing of the manuscript. The verification procedures included comparison with primary references, independent implementations of derivatives, numerical precision checks, and consistency checks between source files and results, as documented in the appendix and accompanying artifacts. The human authors are responsible for the mathematical claims, experimental conclusions, and submitted manuscript.

References

Christopher Aicher, Nicholas J. Foti, and Emily B. Fox. Adaptively truncating backpropagation through time to control gradient bias. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 799–808. PMLR, 2020. URL https://proceedings.mlr.press/v115/aicher20a.html.

Devansh Arpit, Bhargav Kanuparthi, Giancarlo Kerg, Nan Rosemary Ke, Ioannis Mitliagkas, and Yoshua Bengio. h-detach: Modifying the LSTM gradient towards better optimization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.03023v2.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1909.01377.

Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL https://jmlr.org/papers/v18/17-468.html.

Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. One-step differentiation of iterative algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77089–77103, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f3716db40060004d0629d4051b2c57ab-Abstract-Conference.html.

Raoul Bott. Nondegenerate critical manifolds. *Annals of Mathematics*, 60(2):248–261, 1954. doi: 10.2307/1969631. URL https://www.jstor.org/stable/1969631.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1807.03819.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. JFB: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6648–6656, 2022. doi: 10.1609/aaai.v36i6.20619. URL https://ojs.aaai.org/index.php/AAAI/article/view/20619.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15130–15152, 2024. URL https://proceedings.mlr.press/v235/gatmiry24b.html.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach, 2025. URL https://arxiv.org/abs/2502.05171.

Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2111.05177.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. 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/giannou23a.html.

Zixuan Gong, Yong Liu, and Jiaye Teng. What makes looped transformers perform better than non-recursive ones (provably), 2025. URL https://arxiv.org/abs/2510.10089.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758. PMLR, 2020. URL https://proceedings.mlr.press/v119/grazzi20a.html.

Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation, 2026. URL https://arxiv.org/abs/2602.11451.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks, 2025. URL https://arxiv.org/abs/2510.04871.

Georgios Kaissis, David Mildenberger, Juan Felipe Gomez, Martin J. Menten, and Eleni Triantafillou. Step-resolved data attribution for looped transformers, 2026. URL https://arxiv.org/abs/2602.10097v1. Version 1.

Gunn Kim. Dynamical phase selection controls compute scaling in looped transformers, 2026. URL https://arxiv.org/abs/2608.26556.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2013. doi: 10.1007/978-1-4419-9982-5. URL https://doi.org/10.1007/978-1-4419-9982-5.

Shuzhen Li, Yifan Zhang, Jiacheng Guo, Quanquan Gu, and Mengdi Wang. DeepLoop: Depth scaling for looped transformers, 2026. URL https://arxiv.org/abs/2607.13491v2. Version 2.

W. Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953. doi: 10.1090/S0002-9939-1953-0054846-3. URL https://doi.org/10.1090/S0002-9939-1953-0054846-3.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.10369.

Ramon E. Moore, R. Baker Kearfott, and Michael J. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, 2009. ISBN 978-0-89871-669-6. doi: 10.1137/1.9780898717716. URL https://doi.org/10.1137/1.9780898717716.

Sajad Movahedi, Vera Milovanović, Shlomo Libo Feigin, Alexander Theus, Thomas Hofmann, Valentina Boeva, T. Konstantin Rusch, and Antonio Orvieto. Fixed-point reasoners: Stable and adaptive deep looped transformers, 2026. URL https://arxiv.org/abs/2606.18206.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006. doi: 10.1007/978-0-387-40065-5. URL https://doi.org/10.1007/978-0-387-40065-5.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models, 2026. URL https://arxiv.org/abs/2604.12946.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.17416.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1723–1732, 2019. URL https://proceedings.mlr.press/v89/shaban19a.html.

Ayhan Suleymanzade, Chanhyuk Lee, Floor Eijkelboom, Nicholas M. Boffi, İsmail İlkan Ceylan, and Jinwoo Kim. Thinking with looped flows, 2026. URL https://arxiv.org/abs/2609.11801.

Corentin Tallec and Yann Ollivier. Unbiasing truncated backpropagation through time, 2017. URL https://arxiv.org/abs/1705.08209.

Ivan Viakhirev, Kirill Borodin, Amirah Almutairi, Serguei Barannikov, Maxim Abramov, and Grach Mkrtchian. Think shallow, solve deep: Controlling recurrent dynamics for reliable test-time depth, 2026. URL https://arxiv.org/abs/2608.18222.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL https://arxiv.org/abs/2506.21734.

Guan Wang, Changling Liu, Chenyu Wang, Cai Zhou, Yuhao Sun, Yifei Wu, Shuai Zhen, Luca Scimeca, and Yasin Abbasi Yadkori. HRM-Text: Efficient pretraining beyond scaling, 2026. URL https://arxiv.org/abs/2605.20613v1. Version 1.

Ronald J. Williams and Jing Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 2(4):490–501, 1990. doi: 10.1162/neco.1990.2.4.490. URL https://doi.org/10.1162/neco.1990.2.4.490.

Liu Yang, Kangwook Lee, Robert D. Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b8402301e7f06bdc97a31bfaa653dc32-Paper-Conference.pdf.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models, 2025. URL https://arxiv.org/abs/2510.25741.

Appendices

A Proofs for the Normalized Model

All constants in this appendix may depend on the fixed task, C, and the chosen compact manifold patch, but not on m. Gradients are always taken with respect to the eight physical entries of (A, B) unless a coordinate gradient is explicitly specified. For the manifold $\mathcal{M}$ in (5), fix compact patches $\mathcal{M}_0 \Subset \mathcal{M}_1 \Subset \mathcal{M}$ and a neighborhood U of $\mathcal{M}_0$ small enough that the nearest-point projection $\pi_{\mathcal{M}_1}$ is unique (Lee, 2013, Theorem 6.24). Each $\theta \in U$ then has the unique decomposition $\theta = \bar{\theta} + n$, with $\bar{\theta} = \pi_{\mathcal{M}_1}(\theta)$ and n in the five-dimensional normal space at $\bar{\theta}$.

A.1 Data conditions and an explicit task

The data-moment assumption is stated in Section 2.3. This appendix constructs an explicit uniform input distribution satisfying it and verifies the required moments.

For a uniform input distribution on $[-a, a]$, the critical moment $I(a) = \mathbb{E}[r \sin 2X]$ is

$$I(a) = \frac{1}{2}\{\operatorname{sinc}((1+\kappa)a) - \operatorname{sinc}((3-\kappa)a)\}, \qquad \operatorname{sinc}(u) = \frac{\sin u}{u}. \tag{14}$$

We seek a support endpoint a_0 such that $I(a_0) = 0$. The sinc function has value one at zero by continuous extension. The formula follows from $2 \sin u \sin v = \cos(u - v) - \cos(u + v)$ and $\mathbb{E}[\cos(tX)] = \operatorname{sinc}(ta)$ for uniform X on $[-a, a]$. The same product-to-sum identities give

$$\begin{aligned}
J &= \tfrac{1}{2}\{\operatorname{sinc}((3+\kappa)a) - \operatorname{sinc}((5-\kappa)a)\},\\
K &= \tfrac{1}{2}\operatorname{sinc}((1-\kappa)a) - \tfrac{1}{4}\{\operatorname{sinc}((3+\kappa)a) + \operatorname{sinc}((5-\kappa)a)\},\\
(M_2)_{11} &= \operatorname{sinc}((1-\kappa)a),\\
(M_2)_{12} &= \tfrac{1}{2}\{\operatorname{sinc}((1+\kappa)a) + \operatorname{sinc}((3-\kappa)a)\},\\
(M_2)_{22} &= \tfrac{1}{2}\operatorname{sinc}((1-\kappa)a) + \tfrac{1}{4}\{\operatorname{sinc}((3+\kappa)a) + \operatorname{sinc}((5-\kappa)a)\} - J,\\
B_0 &= a\{\operatorname{sinc}'((1+\kappa)a) - \operatorname{sinc}'((3-\kappa)a)\}.
\end{aligned}$$

For $\kappa = 1/2$, use the rational decimal endpoints $a_- = 2.30052398302186$ and $a_+ = 2.30052398302189$. Outward interval evaluation (Moore et al., 2009) gives

$$\begin{aligned}
I(a_-) &\in [1.17, 1.18]\, 10^{-15},\\
I(a_+) &\in [-1.066, -1.065]\, 10^{-14},\\
I'([a_-, a_+]) &\subset [-0.394352951739882, -0.394352951739849],\\
B_0 &\in [-0.956281230254165, -0.956281230254077].
\end{aligned}$$

The intermediate value theorem and the strictly negative derivative give a unique root in this bracket. Evaluating the remaining expressions on the whole bracket gives, rounded outward here for display,

$$\begin{aligned}
J &\in [0.09953298141437, 0.09953298141440],\\
K + J &\in [0.48522328439509, 0.48522328439514],\\
(M_2)_{11} &\in [0.79362000649621, 0.79362000649626],\\
\det M_2 &\in [0.23697409606998, 0.23697409607004].
\end{aligned}$$

The first leading principal minor and the determinant are positive, so the Sylvester criterion gives $M_2 \succ 0$. Every enclosure is evaluated on the same root bracket, so the intermediate value argument and all strict moment inequalities select one common task. The supplied artifact contains the full moment enclosures and a separately recomputed 70-digit interval certificate for opposite endpoint signs, a negative root derivative, nonzero B_0, and the four positive-moment conditions. The certificate evaluates the displayed trigonometric expressions with outward interval arithmetic, propagating the full root enclosure into each moment bound.

A.2 THE CRITICAL SET AND THE EXPLICIT GRADIENT COEFFICIENT

The manifold in (5) has two scale directions and one rotation direction. Its tangent space has dimension three, and its normal space has dimension five. The following result gives the Hessian and coefficient used in Proposition 3.1. A critical manifold is Morse–Bott when the Hessian kernel equals its tangent space (Bott, 1954).

Proposition A.1 (Hessian and finite-depth coefficient)**.** *Under Assumption 2.1, $H(\bar\theta) = \nabla^2 F_\infty(\bar\theta)$ has four positive, one negative, and three zero eigenvalues, with $\ker H(\bar\theta) = T_{\bar\theta}\mathcal{M}$. At $\bar\theta = (\alpha \mathsf{R}_\varphi, \beta \mathsf{R}_{-\varphi})$, the coefficient in (6) is*

$$f_{\mathrm{H}}(\bar\theta) = 0, \quad f_{\mathrm{L}}(\bar\theta) = -\frac{B_0}{4\beta}\mathsf{R}_{-\varphi}(\sin(2\varphi)\mathsf{T} + \cos(2\varphi)\mathsf{S}), \tag{15}$$

where $\mathsf{T} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathsf{S} = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. This vector is nonzero and normal to $\mathcal{M}$. The expansions in (6) hold uniformly on the compact patches for $\|n\| \le K_0 s$, for each fixed K_0.

The omitted contribution $b_m = g_{\mathrm{F},m} - g_{\mathrm{T},m}$ therefore satisfies

$$\begin{aligned} b_m(\bar\theta + n) &= (\mathrm{I} - P)H(\bar\theta)n + sf(\bar\theta) + O(p), \\ g_{\mathrm{T},m}(\bar\theta + n) - P g_{\mathrm{F},m}(\bar\theta + n) &= -sPf(\bar\theta) + O(p). \end{aligned} \tag{16}$$

The term $(\mathrm{I} - P)H(\bar\theta)n$ accounts for the number of retained groups. The second line isolates the signed finite-depth discrepancy after that scaling. The next three subsections establish uniform smoothness, compute the Hessian, and derive the coefficient $f(\bar\theta)$.

A.3 ANGULAR REPRESENTATION AND UNIFORM SMOOTHNESS

For a positive-determinant matrix W near a scaled rotation, let $\phi_W(t) = \arg(We(t))$ denote a continuous local lift. Its derivatives are

$$\phi'_W(t) = \frac{\det W}{\|We(t)\|^2}, \qquad D_W\phi_W(t)[Z] = \frac{\langle \mathsf{J}We(t), Ze(t)\rangle}{\|We(t)\|^2}, \quad \mathsf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{17}$$

For the first identity, differentiate $\arg v$ as $D\arg(v)[z] = \langle \mathsf{J}v, z\rangle / \|v\|^2$ and use $\langle \mathsf{J}We(t), W\mathsf{J}e(t)\rangle = \det W$; the second identity follows by holding t fixed. If $|u - v| < \pi$, the identity $e(u) + e(v) = 2\cos((u-v)/2)e((u+v)/2)$ gives $N(e(u)+e(v)) = e((u+v)/2)$. Each low visit is therefore $l \mapsto \phi_B((l+h)/2)$ in angle coordinates. At $B = \beta\mathsf{R}_{-\varphi}$, its derivative in l is exactly $1/2$. For parameters at distance d from the conformal patch, smoothness gives the uniform bound $q(d) = 1/2 + L_0 d < 1$ after reducing the fixed neighborhood.

The following argument establishes a common angular branch for every visit, including those in arbitrarily long low groups. Choose G with $a_0 < G < \pi$. On a sufficiently small tube around the patch, write $\phi_A(t) = t + \varphi + \eta_A(t)$ and $\phi_B(t) = t - \varphi + \eta_B(t)$, where $\sup_t(|\eta_A(t)| + |\eta_B(t)|) \le Kd$. Reduce the rotation patch and the tube so that $|\varphi| + Kd \le G/2$. If the current angle gap satisfies $|l - h| \le G$, a low update gives $|l^+ - h| \le |l - h|/2 + |\varphi| + Kd \le G$; a high update gives the same bound for $|l - h^+|$. The initial gap is at most a_0, so induction proves the bound for every visit and every m. In particular, $\|e(l) + e(h)\| \ge 2\cos(G/2) > 0$. Matrix singular values also have a common positive lower bound on the compact positive-scale patch. All normalization denominators, and hence all one-visit derivatives through order three, are uniformly controlled. Periodicity handles the choice of angle lift; the number of high updates remains fixed.

We next give the derivative recursions used below. In a conditional low group, include the physical parameters and the incoming high state in ζ, and write $v_{j+1} = T(v_j, \zeta)$. Set $A_j = D_v T(v_j, \zeta)$, $B_j = D_\zeta T(v_j, \zeta)$, $J_j = D_\zeta v_j$, $U_j = D^2_\zeta v_j$, and $V_j = D^3_\zeta v_j$. For a parameter direction a, put $\mathcal{K}_j a = (J_j a, a)$. The chain rule gives

$$\begin{aligned} J_{j+1} &= A_j J_j + B_j, \\ U_{j+1}[a,b] &= A_j U_j[a,b] + D^2 T[\mathcal{K}_j a, \mathcal{K}_j b], \\ V_{j+1}[a,b,c] &= A_j V_j[a,b,c] + D^3 T[\mathcal{K}_j a, \mathcal{K}_j b, \mathcal{K}_j c] \\ &\quad + D^2 T[(U_j[a,b],0), \mathcal{K}_j c] + D^2 T[(U_j[a,c],0), \mathcal{K}_j b] + D^2 T[(U_j[b,c],0), \mathcal{K}_j a]. \end{aligned}$$

All derivatives of T here are evaluated at (v_j, ζ). If $\|A_j\| \le q_* < 1$ and K_k bounds $\|D^k T\|$, these identities imply

$$\begin{aligned}
\|J_{j+1}\| &\le q_*\|J_j\| + K_1,\\
\|U_{j+1}\| &\le q_*\|U_j\| + K_2(1+\|J_j\|)^2,\\
\|V_{j+1}\| &\le q_*\|V_j\| + K_3(1+\|J_j\|)^3 + 3K_2(1+\|J_j\|)\|U_j\|.
\end{aligned}$$

Iterating the first inequality gives $\sup_j \|J_j\| \le \|J_0\| + K_1/(1-q_*)$. Insert that bound into the second inequality and sum its geometric series; then insert both resulting bounds into the third. The resulting constants are independent of the low-group length. Initial-state dependence is included in the bounded initial derivatives. Composing a fixed number C of groups and high updates preserves these bounds.

The same argument applies to the retained sensitivity S_j used by the truncated gradient. Between detachments, $S_{j+1} = A_j S_j + B_j$, with the appropriate retained input sensitivities included in the forcing. A detachment sets the corresponding sensitivity, and all its derivatives, to zero. Its true parameter derivatives otherwise satisfy

$$\begin{aligned}
DS_{j+1} &= A_j DS_j + (DA_j)S_j + DB_j,\\
D^2S_{j+1}[a,b] &= A_j D^2S_j[a,b] + (DA_j[a])DS_j[b] + (DA_j[b])DS_j[a]\\
&\quad + (D^2A_j[a,b])S_j + D^2B_j[a,b].
\end{aligned}$$

The bounds on the forward derivatives control each forcing term. This proves uniform bounds on the first and second derivatives of both gradient fields, including their dependence on the recomputed forward states.

The same recursions give the derivative-tail estimate by comparing each iterate and its parameter derivatives with their conditional-equilibrium counterparts. Let $u_*(\zeta)$ be the conditional fixed point and define

$$e_j^{(0)} = \|v_j - u_*\|, \qquad e_j^{(1)} = \|D_\zeta v_j - D_\zeta u_*\|, \qquad e_j^{(2)} = \|D_\zeta^2 v_j - D_\zeta^2 u_*\|.$$

Contraction, bounded derivatives, and subtraction of the fixed-point recursions give

$$e_j^{(0)} \le q^j e_0^{(0)}, \qquad e_{j+1}^{(1)} \le q e_j^{(1)} + K e_j^{(0)}, \qquad e_{j+1}^{(2)} \le q e_j^{(2)} + K(e_j^{(0)} + e_j^{(1)}),$$

where $q = q(d)$ is common to the group. Thus $e_j^{(1)} \le q^j e_0^{(1)} + K e_0^{(0)} j q^{j-1}$, and a second summation gives $e_j^{(2)} = O((j+1)^2 q^{j-2})$. Implicit fixed-point derivatives are bounded because $\|(\mathrm{I} - D_v T)^{-1}\| \le (1-q_*)^{-1}$. At a group boundary, composition only multiplies these errors by bounded derivatives and adds the preceding state and sensitivity errors. There are only C boundaries. The same subtraction applies to retained sensitivities, including the zero reset, and yields on the fixed tube

$$\|g_{\mathrm{F},m} - g_{\mathrm{F},\infty}\| + \|g_{\mathrm{T},m} - g_{\mathrm{T},\infty}\| \le L_1 m q(d)^m. \tag{18}$$

Constants absorb one factor q^{-1}, since q is uniformly bounded away from zero. At points on $\mathcal{M}$, where the propagation factor is exactly $1/2$, value, first-, and second-derivative tails have orders p, mp, and $m^2 p$. These reference-point rates give the coefficients of the normal Taylor expansion. In the surrounding fixed neighborhood, the position-dependent rate $q(d)$ controls the corresponding tails and links the local expansion to the outer alignment bound.

A.4 The structural Hessian and truncated-field derivative

Together with J from (17), use the trace-free matrices

$$\mathsf{T} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathsf{S} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

They are orthogonal in the Frobenius inner product, each of squared norm two. For $U = u_R\mathsf{J} + u_T\mathsf{T} + u_S\mathsf{S}$, the angular field induced by $\exp(U)$ is $v_U(t) = u_R + u_T\cos 2t + u_S \sin 2t$. The matrix exponential acts through the flow of this field, so its angle map has expansion $t + v_U(t) + \frac{1}{2}v_U'(t)v_U(t) + O(\|U\|^3)$.

First work at $(A, B) = (\mathrm{I}, \mathrm{I})$, writing $A = \exp U$, $B = \exp V$. Solving the low equilibrium equation to second order gives

$$l_*(x) = x + 2v_V(x) + 3v_V'(x)v_V(x) + O(\|V\|^3).$$

Indeed, if $l_* = x+\delta$, its defining equation is $x+\delta = x+\delta/2+v_V(x+\delta/2)+\frac{1}{2}v_V'(x+\delta/2)v_V(x+\delta/2) + O(\|V\|^3)$. Substitution of the first-order solution $\delta = 2v_V$ yields the stated second-order coefficient. Applying H once gives the cycle map

$$Q(x) = x + v_U + v_V + \tfrac{1}{2}v_U'v_U + v_U'v_V + \tfrac{3}{2}v_V'v_V + O(\|(U,V)\|^3).$$

For a map $Q(x) = x + w(x) + z(x) + O(\|(U,V)\|^3)$, with w of degree one and z of degree two in (U, V), induction gives

$$Q^j(x) = x + jw + jz + \tfrac{j(j-1)}{2}w'w + O(\|(U,V)\|^3).$$

Indeed, evaluating w at $Q^j(x)$ adds $jw'w$ at degree two, whereas evaluating z there changes only degree-three terms. The coefficient $j(j-1)/2$ therefore becomes $j(j+1)/2$ at the next step. With $w = v_U + v_V$ and the displayed quadratic part of Q, this gives

$$Q^C(x) = x + Cw + \frac{C^2}{2}w'w + Cv_V'v_V + \frac{C}{2}(v_U'v_V - v_V'v_U) + O(\|(U,V)\|^3). \tag{19}$$

Symmetry and $\mathbb{E}[r\sin 2X] = 0$ imply $\mathbb{E}[r] = \mathbb{E}[r\cos 2X] = \mathbb{E}[r\sin 2X] = 0$. Thus $\mathbb{E}[rv_W] = 0$ for all three angular fields. For scalar fields the bracket coefficient is $[f, g] = fg' - gf'$. Explicitly,

$$[1, \cos 2t] = -2\sin 2t, \qquad [1, \sin 2t] = 2\cos 2t, \qquad [\cos 2t, \sin 2t] = 2.$$

Their brackets therefore lie in the same span. The bracket term in (19) has zero residual expectation. To pass from the angle to the loss, write the final angle as $X + \epsilon$. Then

$$1 - \cos((1-\kappa)X + \epsilon) = 1 - \cos((1-\kappa)X) + r\epsilon + \tfrac{1}{2}c\epsilon^2 + O(|\epsilon|^3).$$

The linear expectation vanishes. Substitution of (19) and cancellation of the bracket term leave the quadratic term

$$\frac{C^2}{2}\mathbb{E}[cw^2 + rw'w] + C\mathbb{E}[rv_V'v_V]. \tag{20}$$

For $w = w_R + w_T\cos 2X + w_S\sin 2X$, direct multiplication gives

$$w'w = -2w_Rw_T\sin 2X + 2w_Rw_S\cos 2X + 2w_Tw_S\cos 4X + (w_S^2 - w_T^2)\sin 4X.$$

The first term has zero residual expectation by the critical moment; the second and third vanish by parity. Therefore

$$\mathbb{E}[rw'w] = J(w_S^2 - w_T^2), \qquad \mathbb{E}[rv_V'v_V] = J(v_S^2 - v_T^2).$$

Likewise the odd cross terms of $\mathbb{E}[cw^2]$ vanish by symmetry. Expanding (20) gives the six-coordinate Hessian

$$\begin{aligned} H_{\mathrm{coord}} &= \begin{pmatrix} C^2M_0 & C^2M_0 \\ C^2M_0 & C^2M_0 + 2CJ\ \mathrm{diag}(0,-1,1) \end{pmatrix}, \\ M_0 &= \begin{pmatrix} \mathbb{E}c & \mathbb{E}[c\cos 2X] & 0 \\ \mathbb{E}[c\cos 2X] & \mathbb{E}[c\cos^2 2X] - J & 0 \\ 0 & 0 & K + J \end{pmatrix}. \end{aligned} \tag{21}$$

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the coefficient vectors of U and V in the displayed basis. The quadratic form is $C^2(\boldsymbol{u}+\boldsymbol{v})^\top M_0(\boldsymbol{u}+\boldsymbol{v}) + 2CJ(v_S^2 - v_T^2)$. The invertible change $(\boldsymbol{u}, \boldsymbol{v}) \mapsto (\boldsymbol{u}+\boldsymbol{v}, \boldsymbol{v})$ is a congruence, so it preserves inertia by the Sylvester law of inertia. Assumption 2.1 makes the first three directions positive; the remaining terms add one positive, one negative, and one zero direction. The two independent matrix-scale directions are also zero, producing exactly four positive, one negative, and three zero physical eigenvalues.

At $\bar\theta = (\alpha\mathsf{R}_\varphi, \beta\mathsf{R}_{-\varphi})$, use $A = \alpha\mathsf{R}_\varphi\exp U$ and $B = \beta\mathsf{R}_{-\varphi}\exp V$. Their angle maps are respectively $\varphi + \phi_{\exp U}$ and $-\varphi + \phi_{\exp V}$, so the same calculation uses phase $X - \varphi$. Set

$$v_\varphi = (1, \cos 2(X-\varphi), \sin 2(X-\varphi))^\top, \quad v_\varphi' = (0, -2\sin 2(X-\varphi), 2\cos 2(X-\varphi))^\top,$$

where the prime denotes differentiation with respect to the angular argument X. Define

$$N_\varphi = \mathbb{E}[r(v'_\varphi v_\varphi^\top + v_\varphi (v'_\varphi)^\top)], \qquad M_\varphi = \mathbb{E}[c v_\varphi v_\varphi^\top] + \tfrac{1}{2} N_\varphi.$$

Then

$$H_{\mathrm{coord}}(\varphi) = \begin{pmatrix} C^2 M_\varphi & C^2 M_\varphi \\ C^2 M_\varphi & C^2 M_\varphi + C N_\varphi \end{pmatrix}. \tag{22}$$

The six columns of the 8×6 matrix $E_{\bar\theta}$ are the physical variations $\alpha \mathsf{R}_\varphi \mathsf{J}, \alpha \mathsf{R}_\varphi \mathsf{T}, \alpha \mathsf{R}_\varphi \mathsf{S}$ in A and $\beta \mathsf{R}_{-\varphi} \mathsf{J}, \beta \mathsf{R}_{-\varphi} \mathsf{T}, \beta \mathsf{R}_{-\varphi} \mathsf{S}$ in B. The second-order coordinate chain rule consists of the Hessian pulled back by $E_{\bar\theta}$ and a coordinate-curvature term multiplied by the physical gradient. At this critical point the latter term vanishes, leaving the following transformation:

$$H(\bar\theta) = E_{\bar\theta} (E_{\bar\theta}^\top E_{\bar\theta})^{-1} H_{\mathrm{coord}}(\varphi) (E_{\bar\theta}^\top E_{\bar\theta})^{-1} E_{\bar\theta}^\top. \tag{23}$$

Scale directions contribute zero, including mixed terms at criticality. Indeed, the coordinate chain rule gives $E_{\bar\theta}^\top H(\bar\theta) E_{\bar\theta} = H_{\mathrm{coord}}$. The six columns span the orthogonal complement of the two radial directions, which are in the kernel; multiplying on both sides by the corresponding dual-frame matrices gives (23). This preserves the physical metric: the factors $(E_{\bar\theta}^\top E_{\bar\theta})^{-1}$ account for the lengths of the six coordinate basis vectors and convert coordinate covectors back into gradients in the original matrix entries.

The gradient vanishes on $\mathcal{M}$, whose tangent space contains the two scaling directions and the counter-rotation direction. Continuity preserves the five nonzero eigenvalues on a sufficiently small compact patch. There can be no additional kernel, so $\ker H(\bar\theta) = T_{\bar\theta}\mathcal{M}$. Let $\sigma_* > 0$ be a uniform lower bound on the smallest singular value of the normal restriction. The Taylor remainder satisfies $\|g_{\mathrm{F},\infty}(\bar\theta + n) - H(\bar\theta) n\| \le L\|n\|^2/2$. Consequently $\|g_{\mathrm{F},\infty}\| \ge \sigma_* \|n\| - L\|n\|^2/2$. For $\|n\| \le \sigma_*/L$ this is at least $\sigma_* \|n\|/2$; the upper bound follows from the uniform upper bound on $\|H\|$ and the same remainder. Since $\|n\| = d$, this proves

$$c_0\, d \le \|g_{\mathrm{F},\infty}(\theta)\| \le C_0\, d, \qquad d = \mathrm{dist}(\theta, \mathcal{M}). \tag{24}$$

This is the local gradient bound associated with Morse–Bott nondegeneracy (Bott, 1954). Its controlling quantity is the smallest singular value of the normal Hessian, so the bound applies to the four positive and one negative normal eigenvalues together.

It remains to justify the truncated-field derivative. In the infinite-low model at (I, I), let $E_1(\delta V)$ denote the separate contribution from the first low group to the derivative of the final angle. Expansion of the cycle map and its propagation through the later $C - 1$ cycles gives

$$\frac{1}{C} D_V Q^C[\delta V] - E_1(\delta V) = \frac{C-1}{2} \{ v'_{\delta V} w - w' v_{\delta V} \} + O(\|(U,V)\|^2). \tag{25}$$

More explicitly, let $a_{\mathrm{loc}}(\delta V)$ be the first-order perturbation of the local cycle source, evaluated at the unperturbed incoming state. It is the same at every cycle. The incoming state at cycle i is $X + (i-1)w + O(\|(U,V)\|^2)$, and the remaining derivative is $1 + (C-i)w' + O(\|(U,V)\|^2)$. Thus

$$E_i(\delta V) = v_{\delta V} + a_{\mathrm{loc}}(\delta V) + (i-1) v'_{\delta V} w + (C-i) w' v_{\delta V} + O(\|(U,V)\|^2 \|\delta V\|).$$

Subtracting the expression for $i = 1$ gives $E_i - E_1 = (i-1)(v'_{\delta V} w - w' v_{\delta V}) + O(\|(U,V)\|^2)$. Averaging over i gives (25); all its remainders are uniform for $\|\delta V\| \le 1$. At the reference point the left side vanishes samplewise, so differentiating the residual adds no first-order discrepancy. The residual average of its displayed first-order term is zero by the bracket closure above. H is exact, while the L-field derivative is r_C times its full derivative. Rotation of the angular basis and blockwise scale changes commute with P, giving

$$D g_{\mathrm{T},\infty}(\bar\theta) = P H(\bar\theta), \qquad \|g_{\mathrm{T},\infty}(\theta) - P g_{\mathrm{F},\infty}(\theta)\| \le L d^2.$$

This proves the Hessian assertions of Proposition A.1 and the first-order compatibility used in Proposition 3.1.

A.5 Computing the omitted finite-depth source

At $\bar\theta = (\alpha \mathsf{R}_\varphi, \beta \mathsf{R}_{-\varphi})$, the low-state angle in the first group after visit j is exactly $X - 2\varphi - 2^{-j}(X - 2\varphi)$. The phase immediately before its matrix action is

$$X - \varphi - 2^{-j}(X - 2\varphi).$$

We compute its downstream weight explicitly. At the conformal pair, one complete cycle is the affine angle map

$$l_c = pl_{c-1} + (1-p)(h_{c-1} - 2\varphi), \qquad h_c = (1-p/2)h_{c-1} + (p/2)l_{c-1} + p\varphi.$$

In the state order (h, l) its Jacobian is

$$J_p = \begin{pmatrix} 1-p/2 & p/2 \\ 1-p & p \end{pmatrix}.$$

For $\delta_c = l_c - h_c + 2\varphi$, subtraction gives $\delta_c = (p/2)\delta_{c-1}$ and $h_c - h_{c-1} = (p/2)\delta_{c-1}$. Since $\delta_0 = -X + 2\varphi$,

$$h_C = X + \delta_0 \frac{p}{2-p}\{1 - (p/2)^C\} = X + O(p).$$

A unit angular injection at low visit j of the first group becomes $2^{j-m}(1/2, 1)^\top$ after its high update. Its final high weight is

$$2^{j-m-1} a_C(p), \qquad a_C(p) = (1,0)J_p^{C-1}(1,2)^\top = 1 + O(p).$$

The last equality follows because C is fixed, the expression is polynomial in p, and $a_C(0) = 1$. Thus the multiplicative factor in this weight is uniform over visits and samples. For the T and S coordinates the source fields are respectively $\cos 2t$ and $\sin 2t$. Expanding at $t = X - \varphi$ gives a first-order phase displacement $-2^{-j}(X - 2\varphi)$. Multiplied by 2^{j-m-1}, this has the same p scale at each visit.

The needed cancellations are

$$\mathbb{E}[r \sin 2X] = \mathbb{E}[r \cos 2X] = 0, \quad \mathbb{E}[Xr \sin 2X] = 0, \quad \mathbb{E}[2Xr \cos 2X] = B_0.$$

Writing out the shifted trigonometric moments gives

$$\mathbb{E}[r(X - 2\varphi) \sin 2(X - \varphi)] = -\tfrac{B_0}{2} \sin 2\varphi,$$
$$\mathbb{E}[r(X - 2\varphi) \cos 2(X - \varphi)] = \tfrac{B_0}{2} \cos 2\varphi.$$

Each visit contributes $-(p/2)\mathbb{E}[r(X - 2\varphi)v'(X - \varphi)]$ to leading order. The derivatives of the two source fields are $-2 \sin 2t$ and $2 \cos 2t$. Their T and S coordinate contributions of this displacement, summed over the first low group, are $-\frac{sB_0}{2} \sin 2\varphi$ and $-\frac{sB_0}{2} \cos 2\varphi$, respectively. The rotation coordinate has no phase-dependent feature and hence no such term. A second-order phase remainder is bounded by $K \sum_{j=1}^{m} 2^{j-m} 4^{-j} = O(p)$. The zeroth-order feature has zero residual expectation. The replacement of the terminal residual by r costs $O(p)$ because its argument differs by $O(p)$ and the sum of visit weights is bounded. Replacing $a_C(p)$ by one costs $O(ps) = O(p)$. For group $c \geq 2$, its incoming conditional gap is $\delta_{c-1} = O(p^{c-1})$. Repeating the weighted phase calculation therefore gives $O(mp^c)$ instead of $O(mp)$; its incoming high angle differs from X by $O(p)$, adding $O(p)$. Direct high-parameter features are also evaluated at phase $X - \varphi + O(p)$, so their expectation is $O(p)$ by the same zero-moment identities. There are only C groups. Hence the order-s source occurs only in the first low group, and deleting that group leaves an $O(p)$ gradient.

Converting the coordinate coefficient to a physical gradient divides by 2β and rotates by $\mathsf{R}_{-\varphi}$, giving precisely (15). Uniformly along the compact patch,

$$g_{\mathrm{F},m}(\bar{\theta}) = sf(\bar{\theta}) + O(p), \qquad g_{\mathrm{T},m}(\bar{\theta}) = O(p). \tag{26}$$

The vector f is orthogonal to both radial directions and to the counter-rotation tangent, hence is normal. The second-derivative tail bound from Appendix A.3 is $O(m^2 p)$. For $\|n\| \leq K_0 s$, Taylor expansion now yields

$$g_{\mathrm{F},m}(\bar{\theta} + n) = sf(\bar{\theta}) + H(\bar{\theta})n + O(p + m^2 p\|n\| + \|n\|^2),$$

and the analogous formula for the truncated field. Since $m^2 p\, s = m^3 p^2 = o(p)$ and $s^2 = o(p)$, this proves (6) with an $O(p)$ remainder.

A.6 Proof of the matching tube theorem

Set $e_m(d) = m(1/2 + L_0 d)^m$, $G = g_{\mathrm{F},\infty}$, and $R = g_{\mathrm{T},\infty} - PG$. Write the finite-depth fields as $g_{\mathrm{F},m} = G + \epsilon_{\mathrm{F}}$ and $g_{\mathrm{T},m} = PG + R + \epsilon_{\mathrm{T}}$. Then $\|R\| \le Kd^2$, $\|\epsilon_j\| \le K e_m(d)$, and

$$D_m = G^\top PG + G^\top R + G^\top \epsilon_{\mathrm{T}} + \epsilon_{\mathrm{F}}^\top PG + \epsilon_{\mathrm{F}}^\top R + \epsilon_{\mathrm{F}}^\top \epsilon_{\mathrm{T}}.$$

The first term is at least $r_C c_0^2 d^2$. Bound each remaining term by the product of its norms, using (24); the $d^2 e_m(d)$ term is absorbed into $d e_m(d)$ on the fixed tube. This yields

$$D_m(\theta) \ge cd^2 - K_1 d^3 - K_2 d\, e_m(d) - K_3 e_m(d)^2. \tag{27}$$

Choose a fixed δ so that $q(\delta) < 1$ and $K_1 d^3 \le (c/4)d^2$ for $d \le \delta$. For the ratio $e_m(d)/d$,

$$\frac{\mathrm{d}}{\mathrm{d}d} \log \frac{e_m(d)}{d} = \frac{mL_0}{1/2 + L_0 d} - \frac{1}{d}.$$

Its unique stationary point, when present, is $1/[2L_0(m-1)]$, and is a minimum. Thus the supremum over $[Ks, \delta]$ is attained at an endpoint. At its lower endpoint,

$$\frac{e_m(Ks)}{Ks} = \frac{1}{K}(1 + 2L_0 Ks)^m \longrightarrow \frac{1}{K},$$

because $ms = m^2 2^{-m} \to 0$. At the upper endpoint $e_m(\delta)/\delta \to 0$. Choosing K large, then m large, makes the last two terms in (27) smaller than $(c/2)d^2$ on the entire interval. This proves the outer statement in one fixed neighborhood.

For the lower statement, evaluate the gradients at the center $\theta_m(\bar\theta)$ defined in (7):

$$g_{\mathrm{F},m}(\theta_m(\bar\theta)) = \tfrac{s}{2} f + O(p), \qquad g_{\mathrm{T},m}(\theta_m(\bar\theta)) = -\tfrac{s}{2} Pf + O(p).$$

Since f is entirely in L, $Pf = r_C f$. Its norm is uniformly bounded below by $|B_0|$ and the upper scale bound on β. Hence the inner product is $-\frac{1}{4} r_C s^2 \|f\|^2 + O(sp + p^2)$, uniformly negative for large m. For the neighborhood of each center, put $z = n/s$ and $\Phi_{\bar\theta}(z) = \langle H(\bar\theta) z + f(\bar\theta), PH(\bar\theta) z\rangle$. At $z_c = -H_N^{-1} f/2$, compactness gives $\Phi_{\bar\theta}(z_c) \le -a$ for a common $a > 0$. The polynomials $\Phi_{\bar\theta}$ have a common Lipschitz constant L_Φ on a fixed ball containing all these centers. Choose $c_{\mathrm{in}} > 0$ with $L_\Phi c_{\mathrm{in}} \le a/4$. The remainder in D_m/s^2 is $O(p/s + (p/s)^2) = O(1/m)$ there, so for large m it is at most $a/4$. Thus the entire normal ball has $D_m \le -(a/2)s^2$; take $c_{\mathrm{asc}} = a/2$. At the center, division of both gradients by s gives nonzero opposite limits $f/2$ and $-r_C f/2$, proving the cosine limit. For the forward claim, $(1/2 + Ks)^m = p(1 + 2Ks)^m \le p \exp(2Kms) = O(p)$ because $ms \to 0$.

The tubular volume element has positive upper and lower bounds on the compact patch. The outer statement places every negative point in a tube with five normal dimensions of radius Ks, giving volume $O(s^5)$. Integrating the inner five-dimensional negative balls over any positive-volume interior three-dimensional manifold patch gives $\Omega(s^5)$. Multiplying by the initialization density bounds proves (8). The three-dimensional base patch has a fixed positive volume, while each five-dimensional normal fiber contributes a factor proportional to s^5. This product structure is what determines the exponent in the lower bound as well as in the outer tube estimate.

A.7 Proof and implementation of the correction

We give the explicit correction and prove Theorem 3.3.

The projection $\Pi_B W = W - \langle W, B\rangle B/\|B\|^2$ removes the component of W parallel to the current low matrix. With $\pi_{\mathcal{M}_1}$ denoting the nearest-point projection onto the larger smooth patch, define

$$\Delta_{\mathrm{H}} = 0, \qquad \Delta_{\mathrm{L}} = -s r_C \Pi_B f_{\mathrm{L}}(\pi_{\mathcal{M}_1}(\theta)), \qquad g_{\mathrm{corr},m} = g_{\mathrm{T},m} - \Delta. \tag{28}$$

The correction uses the data moment B_0, the factor sr_C, and the current matrices. The projection can be computed by a 2×2 eigendecomposition, as shown below.

At a point $\bar\theta$ of the manifold, (28) gives $\Delta = -sPf(\bar\theta)$. The additional projection Π_B changes this by $O(s\,d)$ in a fixed tube, because $f_{\mathrm{L}}(\bar\theta)$ is tangent to the constant-norm sphere at $B(\bar\theta)$.

On the full fixed tube $\|\Delta\| \le Ks$. Combining (27) with $|\langle g_{\mathrm{F},m}, \Delta\rangle| \le Ks(d + e_m(d))$ and the same endpoint argument shows that nonpositive corrected alignment implies $d \le K's$. Inside this smaller tube,

$$g_{\mathrm{T},m} - P g_{\mathrm{F},m} = -sPf(\bar\theta) + O(p), \qquad \Delta = -sPf(\bar\theta) + O(s^2),$$

so $g_{\mathrm{corr},m} = P g_{\mathrm{F},m} + O(p)$. In particular,

$$\langle g_{\mathrm{F},m}, g_{\mathrm{corr},m}\rangle \ge r_C \|g_{\mathrm{F},m}\|^2 - Kp\|g_{\mathrm{F},m}\|.$$

Nonpositive alignment implies $\|g_{\mathrm{F},m}\| = O(p)$. Projecting (6) to the normal space and using the bounded inverse yields $\|n + sH_N^{-1} f\| = O(p)$. This proves (9). The shifted patch has uniformly bounded geometry because the displacement is a smooth $O(s)$ normal field on a compact patch. Its $O(p)$ tube therefore has volume $O(p^5)$.

To justify the projection error used above, write $B = B(\bar\theta) + n_B$. Since $\langle f_{\mathrm{L}}(\bar\theta), B(\bar\theta)\rangle = 0$,

$$\Pi_B f_{\mathrm{L}}(\bar\theta) - f_{\mathrm{L}}(\bar\theta) = -\frac{\langle f_{\mathrm{L}}(\bar\theta), n_B\rangle}{\|B\|^2} B = O(d).$$

The denominator stays uniformly positive. Scale invariance of each normalized matrix action makes both its full and truncated gradients orthogonal to the current matrix. The projection gives the correction the same property.

The projection $\pi_{\mathcal{M}_1}$ has an explicit spectral construction from the current matrices. The following calculation derives it by minimizing Euclidean distance to the positive-scale rotation patch. Form the two row vectors

$$u = \frac{1}{\sqrt{2}}(\langle A, \mathrm{I}\rangle, \langle A, \mathsf{J}\rangle), \qquad v = \frac{1}{\sqrt{2}}(\langle B, \mathrm{I}\rangle, -\langle B, \mathsf{J}\rangle),$$

and let $d_\varphi = (\cos\varphi, \sin\varphi)$ be the leading unit eigenvector of $u^\top u + v^\top v$, oriented so its first coordinate is positive. In the local positive-scale chart, set $\alpha = u d_\varphi^\top/\sqrt{2}$ and $\beta = v d_\varphi^\top/\sqrt{2}$. The resulting pair $(\alpha \mathsf{R}_\varphi, \beta \mathsf{R}_{-\varphi})$ is the nearest point. To see this, the squared distance is

$$\|A\|^2 + \|B\|^2 + 2\alpha^2 - 2\alpha\langle A, \mathsf{R}_\varphi\rangle + 2\beta^2 - 2\beta\langle B, \mathsf{R}_{-\varphi}\rangle.$$

For fixed φ, its derivatives in the scales vanish at $\alpha = \langle A, \mathsf{R}_\varphi\rangle/2$ and $\beta = \langle B, \mathsf{R}_{-\varphi}\rangle/2$, which are the formulas above. Substituting these minimizers leaves the constant $\|A\|^2 + \|B\|^2$ minus $(u d_\varphi^\top)^2 + (v d_\varphi^\top)^2$. Hence the remaining minimization maximizes $(u d_\varphi^\top)^2 + (v d_\varphi^\top)^2$ over unit d_φ. The eigenvalue gap and positive scales persist in a sufficiently small chart, giving a smooth unique projection. The correction is computed from B_0, C, m and this projection. The inverse normal Hessian serves a separate analytic role: it maps the signed gradient displacement to the parameter displacement defining the shifted set $\mathcal{S}_m$.

A.8 Stationary-point displacement

For symmetric data and diagonal initial A and B, reflection symmetry makes all off-diagonal gradients zero without freezing those parameters. On the same-norm diagonal submanifold, write

$$A(a_{\mathrm{H}}) = \sqrt{2}\ \mathrm{diag}(\cos(\pi/4 + a_{\mathrm{H}}), \sin(\pi/4 + a_{\mathrm{H}})),$$

and similarly for B. The physical squared tangent norm is two, so matrix gradient flow has angular velocity minus one half of the coordinate gradient. At the identity the symmetric Hessian block is

$$H_S = C^2(K + J)\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + 2CJ\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We give both the existence argument and the loss calculation. For each field $g_{j,m}$ with $j \in \{\mathrm{F}, \mathrm{T}, \mathrm{corr}\}$, let

$$G_{j,m}(a) = \big(\langle g_{j,m,A}, A'(a_{\mathrm{H}})\rangle, \langle g_{j,m,B}, B'(a_{\mathrm{L}})\rangle\big)^\top, \qquad a = (a_{\mathrm{H}}, a_{\mathrm{L}}).$$

Reflection symmetry removes the off-diagonal components, and scale invariance removes radial components. Therefore $G_{j,m} = 0$ is equivalent to a zero of the full physical field at these points. At the origin, the limiting Jacobians are $J_{\mathrm{F}} = H_S$ and $J_{\mathrm{T}} = \mathrm{diag}(1, r_C) H_S$. Both are invertible because

$\det H_S = 2C^3J(K+J) > 0$. The corrected field has limiting Jacobian $J_{\mathrm{corr}} = J_{\mathrm{T}}$. Uniform C^1 convergence and the derivative bounds established above allow a radius $r > 0$, independent of large m, such that

$$\sup_{\|a\| \le r} \|\mathrm{I} - J_j^{-1} DG_{j,m}(a)\| \le \tfrac{1}{2}.$$

Since $G_{j,m}(0) \to 0$, the map $a \mapsto a - J_j^{-1} G_{j,m}(a)$ maps this ball to itself for large m and is a contraction. It has a unique zero there, with $\|a_j^*\| \le 2\|J_j^{-1}\|\|G_{j,m}(0)\|$. The corresponding matrix pair is $\theta_j^* = (A(a_{j,\mathrm{H}}^*), B(a_{j,\mathrm{L}}^*))$.

The center expansion in coordinate covectors is

$$G_{\mathrm{F},m}(0) = \begin{pmatrix} 0 \\ -sB_0/2 \end{pmatrix} + O(p), \qquad G_{\mathrm{T},m}(0) = O(p).$$

For $\|a\| = O(s)$, the Taylor error beyond $G_{\mathrm{F},m}(0) + H_S a$ is $O(m^2 p\|a\| + \|a\|^2) = O(p)$. Solving the two rows gives $a_{\mathrm{H}} + a_{\mathrm{L}} = O(p)$ and $2CJ a_{\mathrm{L}} = sB_0/2 + O(p)$, hence

$$a_{\mathrm{F,L}}^* = \frac{sB_0}{4CJ} + O(p), \qquad a_{\mathrm{F,H}}^* = -\frac{sB_0}{4CJ} + O(p), \qquad a_{\mathrm{T}}^* = O(p).$$

At the full zero, the corrected field is $O(p)$ by $g_{\mathrm{corr},m} = P g_{\mathrm{F},m} + O(p)$. Applying the same contraction argument in a ball centered there, with radius a sufficiently large multiple of p, gives $a_{\mathrm{corr}}^* - a_{\mathrm{F}}^* = O(p)$. The physical metric satisfies $\|g_{j,m}\|^2 = \|G_{j,m}\|^2/2$ on this submanifold. Substitution of $a_{\mathrm{T}}^* = O(p)$ therefore gives the first gradient estimate below; bounded derivatives give the second.

For the loss difference, uniform Taylor bounds give

$$F_m(a) = F_m(0) + G_{\mathrm{F},m}(0)^\top a + \tfrac{1}{2} a^\top H_S a + O(m^2 p\|a\|^2 + \|a\|^3).$$

The leading linear coefficient is $(0, -sB_0/2)^\top$, and $(H_S^{-1})_{\mathrm{LL}} = 1/(2CJ)$. The decrease at the quadratic minimizer is $\frac{1}{2}(sB_0/2)^2/(2CJ) = s^2 B_0^2/(16CJ)$. Replacing that minimizer and the zero coordinates by a_{F}^* and a_{T}^* costs $O(sp + p^2 + m^2 p s^2 + s^3) = O(mp^2)$. At a_{F}^* the first derivative is zero, so the $O(p)$ corrected displacement changes the loss by $O(p^2)$. We have proved

$$\|g_{\mathrm{F},m}(\theta_{\mathrm{T}}^*)\| = \frac{|B_0|}{2\sqrt{2}} s + O(p), \qquad \|g_{\mathrm{F},m}(\theta_{\mathrm{corr}}^*)\| = O(p),$$

$$F_m(\theta_{\mathrm{T}}^*) - F_m(\theta_{\mathrm{F}}^*) = \frac{B_0^2}{16CJ} s^2 + O(mp^2), \qquad F_m(\theta_{\mathrm{corr}}^*) - F_m(\theta_{\mathrm{F}}^*) = O(p^2). \tag{29}$$

The full-loss Hessian at these points retains a negative transverse eigenvalue in the full parameter space, by (21) and continuity. Equation (29) describes their displacements and loss differences on the symmetry-invariant submanifold.

B Gradient Decomposition, Interventions, and Finite-Depth Certificates

B.1 Loss change under a truncated update

At each parameter value, we recompute the full forward pass before applying the mask. This procedure defines the vector field $g_{\mathrm{T},m}(\theta)$, which need not be the gradient of a scalar objective. When we differentiate this field, we include its dependence on the recomputed forward states and keep the mask fixed. Every loss evaluation after a parameter update also recomputes the complete forward pass.

For any nonzero update vector d, consider the normalized step from Section 2. Local Lipschitz continuity of ∇F_m gives

$$F_m\left(\theta - \eta \frac{d}{\|d\|}\right) = F_m(\theta) - \eta \frac{\langle g_{\mathrm{F},m}, d\rangle}{\|d\|} + O(\eta^2). \tag{30}$$

A strict sign of $\langle g_{\mathrm{F},m}, d\rangle$ determines the loss change for sufficiently small positive steps. Suppose $L_m > 0$ bounds the Lipschitz constant of ∇F_m on the update segment. The Taylor remainder then has magnitude at most $L_m\eta^2/2$. For positive alignment, the descent lemma therefore guarantees a decrease when $0 < \eta < 2\langle g_{\mathrm{F},m}, d\rangle/(L_m\|d\|)$ (Nocedal & Wright, 2006, Chapters 2–3). When the inner product is zero, the first-order term does not determine the loss change. For a nonzero truncated gradient, choosing $d = g_{\mathrm{T},m}$ recovers the sign criterion in (2); the normalization only changes the step length.

The derivative-convergence assumptions stated in the corresponding results also justify $g_{\mathrm{F},\infty} = \nabla F_\infty$.

B.2 Finite-depth gradient identities

The chain rule expresses a shared-parameter gradient as a sum of contributions from its separate uses (Pascanu et al., 2013, Section 2). To define the contribution of cycle i, we assign each low-level group a temporary parameter copy and keep that copy shared across its m updates. We differentiate F_m with respect to the copy for group i, then set all copies equal to θ_{L}. We denote the resulting vector by c_i. Each c_i includes every subsequent path to the loss and the expectation over the data.

For a fixed sample, write the first low-group output as $u_m(\theta) = \mathcal{L}_\theta^m(z; h_0(\theta))$ and the sample loss as $\Psi(\theta, u_m(\theta))$. In $\Psi(\theta, u)$, we treat u as an independent argument and retain all other parameter dependence, including the direct encoder path through h_0. The following chain-rule identity specifies exactly what the mask removes.

Lemma B.1 (Omitted gradient contributions)**.** *Under the setup above,*

$$\begin{aligned} g_{\mathrm{F},m} &= g_{\mathrm{T},m} + b_m, \qquad g_{\mathrm{T},m} = \mathbb{E}[\nabla_\theta \Psi(\theta, u_m)], \\ b_m &= \mathbb{E}[(D_\theta u_m)^\top \nabla_u \Psi(\theta, u_m)]. \end{aligned} \tag{31}$$

For $r_C = (C-1)/C$, the low-parameter blocks satisfy

$$g_{\mathrm{T},m,\mathrm{L}} - r_C g_{\mathrm{F},m,\mathrm{L}} = \frac{1}{C}\sum_{i=2}^{C}(c_i - c_1). \tag{32}$$

Proof. The chain rule for $\Psi(\theta, u_m(\theta))$ separates the partial derivative with u held fixed from the contribution through u_m. The mask retains the first term and removes the second. We take expectations to obtain (31). The shared low parameters receive the sum of the group contributions, so $g_{\mathrm{F},m,\mathrm{L}} = \sum_{i=1}^{C} c_i$ and $g_{\mathrm{T},m,\mathrm{L}} = \sum_{i=2}^{C} c_i$. Hence

$$g_{\mathrm{T},m,\mathrm{L}} - r_C g_{\mathrm{F},m,\mathrm{L}} = \sum_{i=2}^{C} c_i - \frac{C-1}{C}\sum_{i=1}^{C} c_i = \frac{1}{C}\sum_{i=2}^{C}(c_i - c_1).$$

□

If $c_1 = \cdots = c_C$, the truncated low-parameter gradient equals r_C times the full low-parameter gradient. This changes the magnitude of that block and preserves its direction whenever it is nonzero. When cycle contributions differ, (32) gives the signed deviation from this scaling. For $C = 2$, the full low-parameter gradient is $c_1 + c_2$ and the retained gradient is c_2. A scalar reweighting stays in $\mathrm{span}(c_2)$, whereas restoration of c_1 can recover a component outside that span. Section 3 computes the omitted contribution in a normalized loop, and Section 4 compares the contributions of different cycles in a broader model class.

B.3 Adjacent parameter sources and the encoder term

We prove Proposition 4.1 by expressing each cycle contribution through its local parameter response and the remaining computation.

At equilibrium, define the high-level transition and its derivatives by

$$\begin{aligned} T(h) &= \mathcal{H}(h, e_\theta(h)), \quad h_i = T(h_{i-1}), \quad V(h) = D_{\theta_{\mathrm{L}}} T(h), \\ A_i &= D_h T(h_{i-1}), \qquad Q_i = A_C \cdots A_{i+1} \quad (Q_C = \mathrm{I}). \end{aligned} \tag{33}$$

The matrix $V(h)$ is the partial low-parameter derivative with the input h held fixed. The matrix A_i propagates an incoming high-state variation through cycle i, and Q_i carries the output variation from that cycle to h_C. Let $\lambda = \nabla_{h_C}[\ell(\mathrm{Head}_\theta(h_C), y)]$ be the terminal loss adjoint. The limiting contribution of cycle i is $c_{i,\infty} = \mathbb{E}[V(h_{i-1})^\top Q_i^\top \lambda]$. The assumptions are a fixed number $C \geq 2$ of shared H/L macro cycles, disjoint parameter blocks, a unique low equilibrium for each visited high state, and local uniform convergence of the iterated conditional low map and its first derivatives to the equilibrium map. In particular, $D_{l_{\mathrm{in}}}\mathcal{L}^m \to 0$, $D_h\mathcal{L}^m \to D_h e_\theta$, and $D_{\theta_{\mathrm{L}}}\mathcal{L}^m \to D_{\theta_{\mathrm{L}}} e_\theta$ on the visited states and parameter neighborhoods. The encoder, high map, head, and loss are continuously differentiable near the visited and limiting states. Population averages satisfy the domination conditions in Section 2. Smooth uniformly contractive maps with bounded local derivatives satisfy these convergence conditions at fixed C (Grazzi et al., 2020; Geng et al., 2021). For piecewise-affine maps, a sufficient local condition is a uniform strict activation margin at the limiting low and high states, together with uniform strict signs along the finite prefixes leading into their invariant cells on the domain under consideration. These conditions give a common smooth cell description for all sufficiently large depths. The finite-depth certificates below use finite prefix, high-gate, and invariant-ball conditions to establish differentiability of a specified finite computation. The high map is differentiated along the fixed number of macro cycles, with its full Jacobians retained in the ordered products Q_i. The encoder precedes all cycles and the readout follows them. Detaching the first low-group output preserves its forward value and sets its backward sensitivity to zero.

Define the sensitivity discrepancy between adjacent cycles by $\Delta_i = V(h_i) - A_{i+1}V(h_{i-1})$. The explicit response formula for the low-block discrepancy in Proposition 4.1 is

$$e_{\mathrm{L}} := g_{\mathrm{T},\infty,\mathrm{L}} - r_C g_{\mathrm{F},\infty,\mathrm{L}} = \frac{1}{C}\sum_{i=1}^{C-1}(C-i)\mathbb{E}\big[\Delta_i^\top Q_{i+1}^\top \lambda\big]. \tag{34}$$

For a trainable encoder $h_0 = \mathrm{Enc}_\theta(x)$, let $J_{\mathrm{E}} = D_{\theta_{\mathrm{E}}}\mathrm{Enc}_\theta(x)$ denote its parameter Jacobian. Let $E_h = D_h e_\theta(h_0)$ and $H_l = D_l\mathcal{H}(h_0, e_\theta(h_0))$ denote the state derivatives along the omitted path. The omitted encoder contribution at equilibrium is $b_{\mathrm{E}} = \mathbb{E}[J_{\mathrm{E}}^\top E_h^\top H_l^\top Q_1^\top \lambda]$. We use the parameter order and matrix P from Section 2. The limiting alignment is given by (12).

To separate the local parameter-derivative contributions, first give cycle i an independent copy $\theta_{\mathrm{L},i}$ of the low parameters, and only tie these copies after differentiating. At a low equilibrium the previous low state is forgotten, so its cycle output depends on the past only through h_{i-1}. The partial derivative injected by $\theta_{\mathrm{L},i}$ is $V(h_{i-1})$, and the remaining high cycles multiply it by Q_i. Thus $c_{i,\infty} = \mathbb{E}[V(h_{i-1})^\top Q_i^\top \lambda]$ is its gradient contribution. Tying parameter copies sums their derivatives, while the first-group mask removes exactly the contribution from the first L group. Therefore $g_{\mathrm{F},\infty,\mathrm{L}} = \sum_{i=1}^{C} c_{i,\infty}$ and $g_{\mathrm{T},\infty,\mathrm{L}} = \sum_{i=2}^{C} c_{i,\infty}$. Since $Q_i = Q_{i+1}A_{i+1}$,

$$c_{i+1,\infty} - c_{i,\infty} = \mathbb{E}[(V(h_i) - A_{i+1}V(h_{i-1}))^\top Q_{i+1}^\top \lambda].$$

Consequently

$$g_{\mathrm{T},\infty,\mathrm{L}} - \frac{C-1}{C} g_{\mathrm{F},\infty,\mathrm{L}} = \frac{1}{C}\sum_{j=2}^{C}(c_{j,\infty} - c_{1,\infty}) = \frac{1}{C}\sum_{i=1}^{C-1}(C-i)(c_{i+1,\infty} - c_{i,\infty}).$$

The second equality follows by writing $c_{j,\infty} - c_{1,\infty} = \sum_{i=1}^{j-1}(c_{i+1,\infty} - c_{i,\infty})$: the ith difference occurs for exactly $C-i$ values of j. This proves (11) and (34) with the state Jacobians in their original composition order. The telescoping argument uses addition and multiplication of the transported parameter derivatives, so it applies to noncommuting and singular state Jacobians alike. Each product propagates its contribution through the subsequent cycles in their original forward order. Taking norms in (34) also gives

$$\|e_{\mathrm{L}}\| \leq \frac{1}{C}\sum_{i=1}^{C-1}(C-i)\mathbb{E}[\|\Delta_i\|\|Q_{i+1}\|\|\lambda\|].$$

For the encoder, take $J_{\mathrm{E}} = D_{\theta_{\mathrm{E}}}\mathrm{Enc}_\theta(x)$ and put $H_h = D_h\mathcal{H}(h_0, e_\theta(h_0))$. Both products below act on an encoder-parameter perturbation and return the first high-state change. The full first-high sensitivity is $(H_h + H_l E_h)J_{\mathrm{E}}$, whereas the truncated one is $H_h J_{\mathrm{E}}$. Multiplication by Q_1 and pairing

with λ give the missing adjoint $J_{\mathrm{E}}^{\top} E_h^{\top} H_l^{\top} Q_1^{\top} \lambda$. H and head have no parameter sources before the detached node and retain their full gradients. Thus, in the parameter order (E, H, L, head),

$$g_{\mathrm{T},\infty} = P g_{\mathrm{F},\infty} + (-b_{\mathrm{E}}, 0, e_{\mathrm{L}}, 0).$$

Taking its inner product with the full gradient proves (12). Comparing the positive quadratic with both signed terms determines the update direction.

Under (13), each Δ_i is zero because $A_{i+1} = D_h T(h_i)$. Equation (34) then gives $e_{\mathrm{L}} = 0$. If $b_{\mathrm{E}} = 0$ as well, (12) reduces to

$$D_\infty = \langle g_{\mathrm{F},\infty}, P g_{\mathrm{F},\infty} \rangle \geq r_C \|g_{\mathrm{F},\infty}\|^2.$$

Since $r_C > 0$, this is strictly positive whenever $g_{\mathrm{F},\infty} \neq 0$, completing the proof of Proposition 4.1.

B.4 Equilibrium-preserving interventions

Proposition B.2 (Solver invariance of asymptotic alignment)**.** *Under the equilibrium assumptions in Section 2, suppose the low map is strictly contractive on a common convex invariant domain. Replace it by $\mathcal{L}_{\theta,\omega}(l; h) = (1 - \omega) l + \omega \mathcal{L}_\theta(l; h)$ for a fixed $\omega \in (0, 1]$. The conditional equilibrium $e_\theta(h)$ and its derivatives $D_h e_\theta$ and $D_{\theta_{\mathrm{L}}} e_\theta$ are independent of ω. The limiting full and truncated gradients, and hence e_{L}, b_{E}, and D_∞, remain unchanged. The finite-depth convergence rate may depend on ω.*

The next statement records the corresponding finite-depth error bound. The fixed-point calculation below proves both results.

Corollary B.3 (Solver relaxation preserves limiting ascent intervals)**.** *Under the equilibrium assumptions in Section 2, replace the low-level map by its Mann-relaxed counterpart (Mann, 1953) $\mathcal{L}_\omega(l, h) = (1 - \omega) l + \omega \mathcal{L}(l, h)$ with $\omega \in (0, 1]$. If $\mathcal{L}$ remains contractive on a convex invariant domain, then $g_{\mathrm{F},\infty}$ and $g_{\mathrm{T},\infty}$ are invariant to ω. At finite depth m, let ϵ_m^ω bound both gradient errors: $\|g_{j,m}^\omega - g_{j,\infty}\| \leq \epsilon_m^\omega$ for $j \in \{\mathrm{F}, \mathrm{T}\}$. Then*

$$|D_m^\omega - D_\infty| \leq \epsilon_m^\omega \big(\|g_{\mathrm{F},\infty}\| + \|g_{\mathrm{T},\infty}\|\big) + (\epsilon_m^\omega)^2. \tag{35}$$

Under the differentiability and sensitivity-convergence assumptions above, if L has contraction constant $q < 1$, the relaxed map has constant at most $q_\omega = 1 - \omega + \omega q < 1$ for $0 < \omega \leq 1$. It has the same equilibrium because $\mathcal{L}_\omega(e, h) = e$ is equivalent to $\omega(\mathcal{L}(e, h) - e) = 0$, and $\omega > 0$. The common closed convex low-state domain is invariant under relaxation because the relaxed output is a convex combination of the current state and its original low update. Thus both iterations remain in the contraction domain and converge to the same equilibrium for each visited high state. The inverse of $\mathrm{I} - D_l \mathcal{L}$ exists by its convergent Neumann series. Implicit differentiation of the fixed-point equation (Grazzi et al., 2020) gives

$$(\mathrm{I} - D_l \mathcal{L}_\omega)^{-1} D_\theta \mathcal{L}_\omega = [\omega(\mathrm{I} - D_l \mathcal{L})]^{-1} \omega D_\theta \mathcal{L} = (\mathrm{I} - D_l \mathcal{L})^{-1} D_\theta \mathcal{L}.$$

The same equality holds for derivatives with respect to the high state. The macro map and all its physical-parameter derivatives are therefore unchanged. Applying the chain rule to the full and truncated compositions proves Proposition B.2 and the invariance statement in Corollary B.3.

Expand the inner product around $(g_{\mathrm{F},\infty}, g_{\mathrm{T},\infty})$ and apply Cauchy–Schwarz to the two linear error terms and their product. This gives (35). On a compact smooth domain with bounded derivatives, the usual sensitivity estimate gives $\epsilon_m^\omega = O(m q_\omega^m)$; the bound also applies to any other valid error estimate. A strict limiting positive or negative margin survives once the right side is smaller than that margin. This is a statement about low refinement at fixed C and mask.

B.5 Exact residual-feature coverage threshold

Fix the depth m, parameters, and inputs, and consider the affine task family $y(a) = y_0 + a y_1$ under squared loss. Both gradients are affine in a, so we write $g_{j,m}(a) = G_j (1, a)^{\top}$ for $j \in \{\mathrm{F}, \mathrm{T}\}$. Partition the parameters into the exactly retained blocks $R = \{H, \mathrm{head}\}$ and the blocks affected by truncation, $B = \{L, E\}$. We multiply the R-block update by $\rho \geq 0$ and write its alignment as $D_{m,\rho}(a) = (1, a)(\rho K_R + S_B)(1, a)^{\top}$, where

$$K_R = G_R^{\top} G_R, \qquad S_B = \mathrm{Sym}(G_{\mathrm{F},B}^{\top} G_{\mathrm{T},B}). \tag{36}$$

Here $G_R = G_{\mathrm{F},R} = G_{\mathrm{T},R}$, and $\mathrm{Sym}(Z) = (Z + Z^{\top})/2$.

Proposition B.4 (Exact reweighting threshold)**.** *If the retained features satisfy the full-rank coverage condition $K_R \succ 0$, any weight*

$$\rho > \rho_* := \max\left\{0,\ -\lambda_{\min}(K_R^{-1/2} S_B K_R^{-1/2})\right\} \tag{37}$$

guarantees strict descent ($D_{m,\rho}(a) > 0$) for all $a \in \mathbb{R}$. If $0 \le \rho < \rho_*$, the alignment quadratic has a negative direction and takes a negative value at some finite task coordinate. Suppose the coefficient of a^2 is positive at ρ_* and the discriminant has a simple zero there. The negative set is then a finite ascent interval, with width $\Theta(\sqrt{\rho_* - \rho})$ as ρ approaches ρ_* from below.

For squared loss and $y(a) = y_0 + a y_1$, the forward prediction and its full or retained Jacobian do not depend on a. If these Jacobians are J_j, then $g_{j,m}(a) = \mathbb{E}[J_j^\top (f_{\theta,m} - y_0)] - a\mathbb{E}[J_j^\top y_1]$, with any output normalization included in the expectation. These are exactly the two columns of G_j. Let $v(a) = (1, a)^\top$. For the nonnegative R-block update weight ρ,

$$D_{m,\rho}(a) = v(a)^\top (\rho K_R + S_B) v(a).$$

The scalar biased-block product equals its symmetric part, since $v^\top Z v = v^\top \operatorname{Sym}(Z) v$. If $K_R \succ 0$, congruence by $K_R^{-1/2}$ shows that this matrix is positive semidefinite exactly when $\rho \ge -\lambda_{\min}(K_R^{-1/2} S_B K_R^{-1/2})$. Restricting to nonnegative weights gives (37). Strict inequality gives positive definiteness. At equality the quadratic is positive semidefinite; zero values correspond exactly to its null vectors. If that null space contains a vector with nonzero first coordinate, rescaling it produces a finite task coordinate with zero alignment. A null vector proportional to $(0, 1)^\top$ instead represents the limiting task direction as $|a| \to \infty$. For $0 \le \rho < \rho_*$, there is a strictly negative vector. If its first coordinate is nonzero it is proportional to some $v(a)$; if it vanishes, write it as $(0, z_2)$ and replace it by (ϵ, z_2). For all sufficiently small nonzero ϵ, continuity preserves its strictly negative quadratic value, and $a = z_2/\epsilon$ is finite. Thus a finite task coordinate has an ascent direction below the threshold. This proves the full family statement, including the case where the original negative vector lies at the boundary of the affine coordinate chart. The signed matrix S_B includes both L and encoder contributions throughout this argument.

Write $D_{m,\rho}(a) = q_2(\rho)a^2 + q_1(\rho)a + q_0(\rho)$. When $q_2(\rho) > 0$ and the roots are distinct, their separation is

$$\frac{\sqrt{q_1(\rho)^2 - 4q_2(\rho)q_0(\rho)}}{q_2(\rho)}.$$

At a closure with a positive quadratic coefficient and a simple discriminant zero, the width is $\Theta(\sqrt{\rho_* - \rho})$. More explicitly, if the discriminant is positive for $\rho < \rho_*$ and vanishes simply at ρ_*, then it equals a positive constant times $\rho_* - \rho + O((\rho_* - \rho)^2)$. Since the quadratic coefficient converges to a positive value, the square root of the discriminant determines the leading rate of the width. If K_R is singular and a finite task has $v(a)^\top K_R v(a) = 0$ together with $v(a)^\top S_B v(a) < 0$, its alignment stays strictly negative at every finite retained-block weight.

B.6 Invariant activation cells and all-parameter sensitivities

The nonlinear residual experiment uses a piecewise-affine recurrent network with leaky-ReLU activation ($\sigma(t) = \max(t, 0.2t)$):

$$\begin{aligned} \mathcal{L}(l, h) &= Wl + Bh + b + U\sigma(Vl + Sh + c), \\ \mathcal{H}(h, l) &= W_{\mathrm{H}} h + B_{\mathrm{H}} l + b_{\mathrm{H}} + U_{\mathrm{H}}\sigma(V_{\mathrm{H}} h + S_{\mathrm{H}} l + c_{\mathrm{H}}). \end{aligned} \tag{38}$$

For each fixed h, the bound $\||W| + |U||V|\|_\infty < 1$ guarantees global contraction of the low map in l. Here absolute values are entrywise, and $\|\cdot\|_\infty$ is the induced matrix infinity norm. In an activation cell with slope matrix $\Gamma = \operatorname{diag}(\gamma_i)$ and $\gamma_i \in \{0.2, 1\}$, the low map is affine in l, with linear part $A = W + U\Gamma V$.

Proposition B.5 (Exact forward rollouts within an activation cell)**.** *Let $u = (\mathrm{I} - A)^{-1}(Bh + b + U\Gamma(Sh + c))$ be the cell equilibrium. If a prefix output l_k satisfies*

$$\|A\|_\infty < 1 \quad \textit{and} \quad s_i(Vu + Sh + c)_i > \|V_i\|_1 \|l_k - u\|_\infty \quad \forall i, \tag{39}$$

where $s_i \in \{-1, 1\}$ is the sign defining the candidate cell and V_i is row i of V, then the closed ℓ_∞ ball centered at u with radius $\|l_k - u\|_\infty$ is forward-invariant. In particular,

$$l_m = u + A^{m-k}(l_k - u), \qquad \forall m \ge k. \tag{40}$$

For any two scalar inputs, the leaky-ReLU secant slope lies in $[0.2, 1]$. Hence, including segments crossing a kink,

$$|\mathcal{L}(l,h) - \mathcal{L}(\widetilde{l},h)| \le (|W| + |U||V|)|l - \widetilde{l}|$$

entrywise. Taking the infinity norm proves the stated global contraction condition over the whole state space, including segments that cross an activation boundary. Within an open cell, every low-state derivative is $W+U\Gamma V$, with diagonal entries of Γ in $\{0.2, 1\}$. In a candidate cell let $A = W+U\Gamma V$, $B_\Gamma = B + U\Gamma S$, $b_\Gamma = b + U\Gamma c$, and $u = (\mathrm{I} - A)^{-1}(B_\Gamma h + b_\Gamma)$. Put $r = \|l_k - u\|_\infty$. For $\|l - u\|_\infty \le r$,

$$s_i(Vl + Sh + c)_i \ge s_i(Vu + Sh + c)_i - \|V_i\|_1 r > 0.$$

Hence the activation pattern of the candidate cell is constant throughout the entire ball, and $\|\mathcal{L}(l,h) - u\|_\infty \le \|A\|_\infty r \le r$ maps it back into itself, including the singleton ball when $r = 0$. Induction proves the exact finite tail. For relaxation, replace A by $(1-\omega)\mathrm{I} + \omega A$ and scale the forcing by ω. Each sample and each H cycle has its own invariant ball and activation pattern. The corresponding affine descriptions therefore assemble into the original nonlinear network, with each description justified on the states reached by its own conditional low group.

Strict prefix and H preactivation signs, and strict invariant-ball margins, persist in a parameter neighborhood. The finite network and the tail formula then agree on that neighborhood, so their physical-parameter derivatives agree. In particular, for $n = m - k$,

$$d\{u + A^n(l_k - u)\} = du + A^n(dl_k - du) + \sum_{j=0}^{n-1} A^j(dA)A^{n-1-j}(l_k - u).$$

This includes parameter-dependent activation-cell matrices, nonzero initial states, encoder contributions, and all cross-channel couplings. After the first low group, the truncated calculation zeroes its sensitivity while retaining its exact value. Subsequent shared parameter injections remain active.

The activation pattern Γ stays fixed in the certified parameter neighborhood, while the matrix factors W, U, V remain trainable. The equilibrium also changes through the affine forcing and incoming high state. Differentiating these quantities gives

$$\begin{aligned} dA &= dW + (dU)\Gamma V + U\Gamma(dV), \\ du &= (\mathrm{I} - A)^{-1}\{(dA)u + (dB_\Gamma)h + B_\Gamma dh + db_\Gamma\}. \end{aligned}$$

The identity for $d(A^n)$ follows by induction: differentiating $A^{n+1} = A^n A$ appends the term $A^n(dA)$ to the preceding sum multiplied by A. Thus every position of dA occurs once, with its original order.

B.7 Rational error and root certificates

The certified mathematical instance is the decimal-rational interpretation of the stored parameters and samples, with rational leak and relaxation. Rational first-order forward-mode differentiation (Baydin et al., 2018) propagates exact values and all 74 parameter derivatives; Gauss–Jordan inversion differentiates via $dA^{-1} = -A^{-1}(dA)A^{-1}$, and matrix powers use product rules. A prefix or high-module activation at a kink violates the strict activation condition required for a differentiable cell certificate. Invariant-ball margins and cell contraction are checked as exact rational inequalities, so a successful check proves the sign margin for the specified instance.

At fixed depth m and block weight ρ, write the exact task quadratic and its frozen binary64 prediction as

$$D_{m,\rho}(a) = q_2 a^2 + q_1 a + q_0, \qquad \widehat{D}_{m,\rho}(a) = \widehat{q}_2 a^2 + \widehat{q}_1 a + \widehat{q}_0.$$

Representing the binary64 coefficients as rationals gives exact coefficient differences. For every real task coordinate,

$$|D_{m,\rho}(a) - \widehat{D}_{m,\rho}(a)| \le |q_2 - \widehat{q}_2|a^2 + |q_1 - \widehat{q}_1||a| + |q_0 - \widehat{q}_0|.$$

The finite tail is exact on a valid cell, so this bound measures the numerical coefficient error in the frozen prediction. For $q_2 > 0$, the rational discriminant decides whether a negative interval exists. When it is positive, the rational vertex $a_* = -q_1/(2q_2)$ satisfies $D_{m,\rho}(a_*) = -(q_1^2 - 4q_2q_0)/(4q_2) < 0$. Choose rational $R > 0$ with $q_2R^2 > (q_1^2 - 4q_2q_0)/(4q_2)$; then $D_{m,\rho}(a_* - R) > 0$ and $D_{m,\rho}(a_* + R) > 0$. Since $D'_{m,\rho}(a) = 2q_2(a - a_*)$, each side contains exactly one root. Bisect

its sign-changing rational bracket and retain the half with opposite endpoint signs; an exactly zero midpoint certifies the root immediately. After n bisections each initial width is reduced by 2^{-n}, yielding disjoint root intervals of width at most 10^{-8}. A negative discriminant certifies a positive minimum for the entire task family. Both the root enclosures and the positive-minimum certificate are computed from the rational finite-response coefficients. Independent unrolled gradients are evaluated subsequently to check the predictions.

The certificate is exact for the specified rational model. After the certificates are saved, an independent reverse-mode implementation differentiates the finite unrolled graph in float64.

C Additional Related Work

Universal Transformers introduced recurrent computation over depth (Dehghani et al., 2019). Subsequent theory characterizes what shared parameters can compute: looped Transformers can implement algorithms (Giannou et al., 2023), and repeated applications can learn or express iterative procedures in context (Yang et al., 2024; Gatmiry et al., 2024); related work on latent reasoning and recurrent language models studies how additional iterations affect predictions and reasoning performance (Saunshi et al., 2025; Geiping et al., 2025; Zhu et al., 2025). The gap between forward refinement and backward contributions also appears in hierarchical and recursive reasoning models: HRM organizes computation into interacting high- and low-level states, and TRM and HRM-Text study recurrent refinement and its training (Wang et al., 2025; Jolicoeur-Martineau, 2025; Wang et al., 2026). The schedule we analyze fixes the number of high-level cycles and varies the number of low-level iterations within each. Its backward mask removes the first low-level group, which specifies exactly which shared-parameter contributions are missing and lets us ask whether additional low iterations change their direction or only their magnitude.

Truncated backpropagation has a long history in recurrent training (Williams & Peng, 1990), and the chain-rule representation as a sum of parameter derivatives transported through later states underlies the analysis of recurrent gradients (Pascanu et al., 2013). Truncation methods differ in which terms of this sum they retain and how they weight them. ARTBP uses random truncation points and compensation factors to obtain an unbiased estimate over the truncation randomness, while its experiments show that fixed truncation can diverge in a worst-case synthetic task (Tallec & Ollivier, 2017). Adaptive TBPTT selects the truncation length to control gradient bias (Aicher et al., 2020), and h-detach modifies the balance between recurrent gradient paths (Arpit et al., 2019). These results establish that omitted contributions can affect gradient accuracy and update direction; we quantify where reversal occurs under a fixed mask and how that region changes with depth.

Uniform contraction of the recurrent state map gives quantitative control of truncation error. Under this notion of stability, Miller & Hardt (2019) show that stable recurrent networks are well approximated by feed-forward models for both inference and gradient training. Our backward mask instead preserves the complete forward trajectory, so the comparison is between two derivatives of the same prediction. Near a critical set the update direction depends on the size and sign of the omitted contribution relative to the local full gradient, and we compute that relation from the loop structure, obtaining the location, matching thickness, initialization probability, and finite-time consequences of the ascent region.

Deep equilibrium models differentiate fixed points (Bai et al., 2019), and analyses of truncated reverse-mode differentiation and hypergradient computation quantify how iterative derivatives approach their limiting values (Shaban et al., 2019; Grazzi et al., 2020). Jacobian-free backpropagation avoids the inverse-Jacobian solve in the backward pass (Fung et al., 2022). Phantom gradients approximate the implicit response using damped unrolling or truncated Neumann series; Geng et al. (2021, Theorem 1) give sufficient conditions for positive alignment with the full gradient, so the negative approximate gradient is a descent direction. One-step differentiation analyzes the derivative obtained from the final iteration and separates forward solution error from the error of omitting earlier derivative responses (Bolte et al., 2023). Its asymptotic accuracy result assumes that the state Jacobian vanishes at the fixed point, a condition satisfied by the superlinearly convergent iterations considered in that work (Bolte et al., 2023, Corollary 2). Together these results distinguish convergence of a solver from convergence of the derivative used to train through it, and our equilibrium calculation applies that distinction to the hierarchical mask above: differentiating the conditional low equilibrium determines the parameter contribution of each cycle, the mask selects which contributions enter the update, and

the cyclewise identity expresses their signed discrepancy after accounting for the retained cycle count. Relaxing a contractive low solver preserves its equilibrium and implicit derivatives, hence also the limiting gradients. Fixed-Point Reasoners connect stable recurrent computation and adaptive stopping to Neumann-series bounds for approximating fixed-point derivatives (Movahedi et al., 2026); our emphasis is the geometry and sign of the error under a specified mask.

DeepLoop analyzes shared-parameter gradient accumulation and residual scaling, including hierarchical loops (Li et al., 2026): one update aggregates gradients across visits, and the updated parameters affect every visit in the next forward computation, so alignment across visits changes the residual scaling in their sufficient stability bound. Step-resolved attribution likewise decomposes training influence across iterations (Kaissis et al., 2026). Our cyclewise decomposition instead compares the contributions retained and removed by a specified mask; their signed difference, with the local loss geometry, determines whether increasing low-level depth confines ascent directions near a critical set or leaves a nonzero limiting discrepancy. Other work studies complementary aspects of recurrent computation: stability and scaling (Prairie et al., 2026), landscape models relating optimization to valley geometry (Gong et al., 2025), the evolution of representations and predictions with depth (Jeddi et al., 2026; Viakhirev et al., 2026; Kim, 2026), and local denoising objectives that coordinate recursive computation (Suleymanzade et al., 2026).

D Experimental Settings

This appendix describes the experimental settings and numerical checks used in Sections 3 and 5. Within each configuration, every comparison uses the same parameters and forward objective.

D.1 Normalized-model experiments

This subsection describes the task construction, geometry grid, and stationary-point protocol for the normalized-model experiment in Section 5.1.

We define $X = a_0 Z$, where Z has density $[1 + \xi(z^2 - 1/3)]/2$ on $[-1, 1]$, $\kappa \in \{1/4, 1/2, 3/4\}$, and $\xi \in \{-0.2, 0, 0.2\}$. These choices define nine tasks. For each task, we solve the critical moment equation once to determine a_0, then keep the distribution and target fixed across all depths. We test $C \in \{2, 4\}$, $m \in \{8, 12, 16, 20, 24\}$, and the manifold locations

$$(\varphi, \alpha, \beta) \in \{(0, 1, 1), (0.025, 0.93, 1.07), (-0.025, 1.07, 0.93)\}.$$

At each location, we draw three normal directions with seed $85213 + 100i + 10C + j$, where i and j index the task and location. We place offsets from the structural center at normalized radii $0, 0.5, 0.9, 1.1, 2$. A second grid uses the fixed radii $0.005, 0.01, 0.02, 0.04$. The two grids contain 4,050 and 3,240 configurations, respectively.

For Figure 3(b), we compute stationary points on the symmetric same-norm diagonal submanifold for all nine tasks, both cycle counts, and all five depths. The solver uses 96-point Gauss–Legendre quadrature with a tolerance of 10^{-10}. We accept a solution if its coordinate-gradient norm is at most 3×10^{-11} and both angular coordinates have magnitude below 0.15. This criterion accepts solutions from 268 of the 270 runs. We recompute the small loss gaps at 60-digit precision.

D.2 Cycle contributions, boundaries, and correction

This subsection describes the protocol for the cycle-contribution and normal-line-boundary experiments reported in Section 3 and the related normalized-model results in Section 5.1.

For the cycle-contribution experiment in Figure 2, we use the nine tasks, two cycle counts, three manifold locations, and five depths defined above. The figure shows the uniform task with $\kappa = 1/2$ at $\bar{\theta} = (\mathrm{I}, \mathrm{I})$. It orders the nine tasks first by increasing κ and then by increasing ξ, so the displayed task appears in row 5.

To measure each low-level-group contribution, we assign a separate copy of B to each group and share that copy across its m applications. All high-level updates share the same matrix A. We evaluate the derivative with respect to each copy at the common value of B. The sum of these derivatives gives the full low-block gradient, while the sum over groups $2, \ldots, C$ gives the truncated low-block

gradient. Across the 270 configurations, we record the first-group contribution, the sum of the later groups, and the high-block contribution.

For the boundary experiment, we set $\bar{\theta} = (\mathrm{I}, \mathrm{I})$ and evaluate

$$\theta = \bar{\theta} - usH_N(\bar{\theta})^{-1} f(\bar{\theta}).$$

We use all nine tasks, $C \in \{2, 4\}$, and $m \in \{12, 16, 20, 24\}$. For each of the 72 configurations, we evaluate 129 values of $u \in [-0.5, 1.5]$ to bracket the sign changes. We then refine each boundary to a tolerance of 10^{-11}. The gradients include all eight entries of (A, B). Every configuration produces two boundaries that enclose a single ascent interval. The measured interval becomes narrower as depth increases.

D.3 Matched updates for the correction ablation

This subsection describes the matched-update protocol for the structural-correction ablation reported in Section 3.

The ablation in Table 1 includes all 54 combinations of task, cycle count, and manifold location from the contribution experiment at $m \in \{12, 16\}$. For each configuration, we set the parameters to the predicted center $\theta_m(\bar{\theta})$ in (7). We compare the truncated gradient $g_{\mathrm{T},m}$, the rescaled gradient $P^{-1}g_{\mathrm{T},m}$, the reversed correction $g_{\mathrm{T},m} + \Delta$, the structural correction $g_{\mathrm{T},m} - \Delta$, and the full gradient $g_{\mathrm{F},m}$. Equation (28) defines Δ. For each direction d, we apply the update $\theta^+ = \theta - \eta d/\|d\|$ and recompute the same finite-depth objective.

We report results at the step norm $\eta = s/12800$. To check that the updates remain in the local regime, we also use the common refinement schedule $s/1600, s/3200, s/6400, s/12800, s/25600$. All 2,700 refined updates have the same sign as their directional derivatives. Halving the reported step also preserves every sign.

D.4 Nonlinear residual model

This subsection describes the protocol and certificates for the nonlinear residual-model experiment in Section 5.2.

The experiment in Figure 4 uses the current-low-state residual model in (38). The model has two-dimensional states, hidden width three, two outputs, a Leaky-ReLU slope of 0.2, two high-level cycles, and 74 nonzero trainable parameters. We create four networks by combining seeds 1703 and 1713 with feedback values $\tau \in \{-0.95, -0.85\}$. The fixed parameter arrays in the artifact package specify each rounded instance.

For each data seed in $\{7101, 7102\}$, we draw eight inputs uniformly from $[-0.08, 0.08]^2$ and nonzero initial low states from $[-0.056, 0.056]^2$. We define two affine target families $y(a) = y_0 + ay_1$:

$$\text{contrast:}\quad y_0 = (1 + 0.3x_1 + 2x_1x_2,\ 0.1x_2 + x_1^2), \qquad y_1 = (0.05 + 0.1x_1,\ 1 + 0.3x_2),$$
$$\text{product:}\quad y_0 = (0.3 + x_1 + 2x_1x_2,\ -0.2 + x_2 + x_1^2), \quad y_1 = (1 + 0.5x_2,\ -0.7 + 0.3x_1).$$

We define the loss as one half of the mean squared error over all samples and output coordinates. We keep the networks and batches fixed across depths.

We evaluate

$$m \in \{6, 8, 10, 12, 14, 16, 20, 24, 32, 48\}, \qquad \omega \in \{0.6, 0.8, 1\}, \qquad \rho \in \{0.8, 1, 1.25\}.$$

The four networks and 16 network–task batches produce 480 forward configurations and 1,440 alignment quadratics. We test prefix lengths in the fixed order $2, 4, 6, 8, 12, 16, 20$, subject to $k \le m - 2$. The invariant-cell condition resolves 312 configurations, which correspond to 936 quadratics. We also evaluate every configuration by independently differentiating the complete unrolled graph.

For the matched-update test, the fixed selection rule chooses the task coordinate that minimizes the positive-curvature quadratic at $0.8\rho_*$. We reuse the same task, parameters, and forward graph at $1.2\rho_*$. At $\omega = 1$ and $m \in \{24, 48\}$, we normalize both directions to 10^{-5} and 5×10^{-6}. This design gives the 128 updates reported in Section 5.2. For the representative settings $(\omega, m) = (1, 12), (1, 24), (0.6, 24)$, exact rational certificates verify an ascent interval at $\rho = 1$ and its removal at $\rho = 1.25$.

D.5 LOOPED-TRANSFORMER INTERVENTIONS

This subsection describes the training and evaluation protocol for the looped-Transformer experiment in Section 5.3.

Each high- and low-level module contains single-head causal attention and a two-layer GELU MLP. The module uses pre-normalization and applies RMSNorm at its output; the RMSNorm layer has no trainable affine parameters. The model has width 64, sequence length 64, an independent embedding and decoder, a fixed nonzero initial low state, and 73,856 trainable parameters. We train with $C = 2$ and $m = 6$, truncate the first low-level group, and update every parameter block.

We split Tiny Shakespeare by character position into 80% training, 10% calibration, and 10% evaluation partitions. The windows and their next-character labels do not overlap across partitions. For seeds 9011–9013, we train each model for 600 AdamW steps with batch size 8, learning rate 10^{-3}, weight decay 0.01, and gradient clipping at norm 1. We evaluate each model at steps 200 and 600, which gives six checkpoints. Eight calibration windows determine the intervention ranking. The evaluation set contains 16 new windows selected with seed 99172.

We use the empirical cross-entropy at $m = 6$ as the reference objective. The comparison includes the truncated gradient, a variant with twice the low-level-block weight, restoration of the omitted first-group low-parameter derivatives, a truncated gradient computed at $m = 12$, the full gradient, and the optimal nonnegative block rescaling. The restoration method keeps the original truncation of the encoder paths. We average each direction over examples before we compute its cosine similarity with the full gradient. We evaluate every method on the same $m = 6$ objective, including the direction computed at $m = 12$. The matched updates have Euclidean norms 10^{-4} and 5×10^{-5}. After each update, we recompute the complete reference forward pass and restore the original parameters. We fix the calibration ranking before evaluation on the held-out windows.