LOOP-μP: RESIDUAL SCALING AND HYPERPARAMETER TRANSFER IN LOOPED TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Looped Transformers have drawn growing attention because looping improves multi-step reasoning and increases the executed depth of a model without adding parameters. They achieve this by reusing a stack of L layers for N loops. However, increasing the loop count N not only destabilizes training but also shifts the best hyperparameters, so they must be searched again for every loop count. Inspired by Depth-μP (Yang et al., 2024a), which multiplies each residual branch by $L^{-1/2}$ for depth L so that a learning rate tuned on a shallow untied network transfers to a deep one, we study how the residual branches and the learning rate should scale with the loop count N. We find that loops cannot simply be treated as extra depth, because the loops share weights. At initialization, distinct layers add nearly uncorrelated increments to the residual stream, so their sum grows as $\sqrt{L}$, whereas a shared layer adds positively correlated increments at every loop, so their sum grows linearly in N. For a single layer looped N times, we show in the infinite-width limit that multiplying the residual branch by $1/N$ and keeping the learning rate independent of N is the only parameterization with both a stable forward pass and maximal residual-stream updates. We then generalize the analysis to a stack of L distinct layers looped N times to obtain a parameterization we call Loop-μP, in which the loop factor $1/N$ combines with the depth factor $L^{-1/2}$. Empirically, we find in language modeling that the base learning rate tuned on the smallest model stays best under Loop-μP for loop counts up to 8 and executed depths up to 384 layers. These results make Loop-μP a practical recipe for scaling looped Transformers.

1 INTRODUCTION

Looped Transformers, which apply the same stack of L layers N times (Dehghani et al., 2019; Lan et al., 2020), have drawn growing attention because they reach an executed depth of NL layers while storing the parameters of only L layers. Looped Transformers can express and learn iterative algorithms by running one step per loop (Saunshi et al., 2025; Gatmiry et al., 2024), and on synthetic tasks they match standard Transformers with far fewer parameters and generalize to longer inputs (Yang et al., 2024b; Fan et al., 2025). In language modeling, Saunshi et al. (2025) and Zhu et al. (2025) observe that downstream accuracy rises as the loop count N grows, most strongly on reasoning tasks. Increasing the loop count N is therefore a natural way to scale a looped model.

However, increasing the loop count N not only destabilizes training but also changes the best learning rate, requiring retuning at every N (Geiping et al., 2025; Zhu et al., 2025; Prairie et al., 2026). For example, Zhu et al. (2025) reduce the loop count after observing loss spikes, and they adopt a conservative learning rate because their compute budget does not allow a learning-rate sweep.

Since a looped model executes $L \times N$ layers, a natural fix is to treat the loops as extra depth. Deep networks solve this problem by introducing depthwise parameterizations, which shrink each residual branch as the depth L grows (Yang et al., 2024a; Dey et al., 2025). Each branch f updates the residual stream as $\boldsymbol{x} \leftarrow \boldsymbol{x} + c\, f(\boldsymbol{x})$, and because a deeper network adds more such updates, these parameterizations set the branch multiplier to $c \propto L^{-\alpha_L}$. Theory supports the whole range $\alpha_L \in [1/2, 1]$, from Depth-μP ($\alpha_L = 1/2$) (Yang et al., 2024a; Bordelon et al., 2024) to CompleteP ($\alpha_L =$

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.

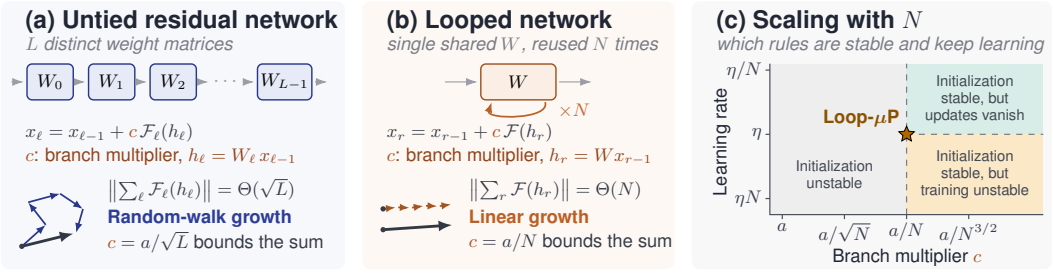


Figure 1: **Weight sharing changes how residual contributions add up, leaving one stable scaling with N. (a)** In an untied network, the L layers contribute nearly independent directions, so residual stream grows as $\Theta(\sqrt{L})$ and the branch multiplier $c = a/\sqrt{L}$ with a constant $a > 0$ keeps it bounded. **(b)** In a looped network, one block is reused N times and its contributions are positively correlated, so residual stream grows as $\Theta(N)$ and c must scale as a/N. **(c)** Each parameterization sets $c = aN^{-\alpha_N}$ and the Adam learning rate $\eta N^{-\gamma_N}$. A branch multiplier that decays more slowly than $1/N$ ($\alpha_N < 1$) lets the residual stream grow with N at initialization (left). For $\alpha_N > 1$, $\gamma_N \geq 0$ makes the updates vanish (upper right), and $\gamma_N < 0$ makes training unstable (lower right). At $\alpha_N = 1$, the same holds for every $\gamma_N > 0$ and $\gamma_N < 0$. The only choice that keeps the residual stream bounded and the updates of order one is therefore $\alpha_N = 1$ and $\gamma_N = 0$, which is Loop-μP.

1) (Dey et al., 2025), and experiments find that both ends transfer the learning rate (Mlodozeniec et al., 2025; Hägele et al., 2026). We ask whether the loop count N admits the same range of exponents.

We answer this question by analyzing a single residual block looped N times. At initialization, the loops share the same weights, so their residual increments $c\,f(\boldsymbol{x}_{r-1})$ are positively correlated, and their sum $\sum_{r=1}^{N} c\,f(\boldsymbol{x}_{r-1})$ grows linearly with N, whereas the increments $c\,f_\ell(\boldsymbol{x}_{\ell-1})$ of L distinct layers are independent and their sum $\sum_{\ell=1}^{L} c\,f_\ell(\boldsymbol{x}_{\ell-1})$ grows only as $\sqrt{L}$ (Figure 1). Figure 2 validates this analysis in a Transformer. While a branch multiplier proportional to $L^{-1/2}$ keeps the residual stream of an untied model of order one, only a branch multiplier proportional to N^{-1} does so for a looped model. Similarly, during training, all N loops receive the same weight update, so the changes it causes in their residual increments are also positively correlated, and the change in the residual stream after one training step grows linearly with N. Therefore, only a branch multiplier of $1/N$ keeps the changes in the residual stream of order one (Figure 3, right). This differs from depth, which admits every exponent in $[1/2, 1]$ (Yang et al., 2024a; Bordelon et al., 2024; Dey et al., 2025).

We formalize this argument in the spirit of the Tensor Programs series (Yang & Hu, 2021; Yang et al., 2024a). We describe a parameterization by how its branch multiplier and learning rate scale with N, and we classify all such parameterizations as the width and then N grow to infinity. Only the branch multiplier $1/N$ with an N-independent learning rate keeps the forward pass stable and the residual-stream updates of order one (Section 3, Figure 1c). We prove this classification for linear networks trained with SGD and extend it to blocks with multiple weight matrices, such as the attention and MLP blocks of a Transformer. For nonlinear activations and Adam, we justify it under additional assumptions (Appendices A, G and I).

Furthermore, we generalize the analysis from looping a single block to looping a stack of L independent layers, which yields a parameterization we call **Loop-μP**. Loop-μP multiplies each residual branch by a factor proportional to $1/(N\sqrt{L})$ and keeps the Adam learning-rate factor $1/\sqrt{L}$, with no factor of N (Appendices B and H). Table 1 summarizes Loop-μP as a recipe for prenorm Transformers. (Section 3.3 and Appendix B.4).

Empirically, we validate Loop-μP by training looped Transformers for language modeling on FineWeb-Edu (Penedo et al., 2024). We find that under Loop-μP, the learning rate tuned on the smallest model stays best for loop counts up to 8 and executed depths up to 384 layers. We also compare Loop-μP with three other loop factors at 12 layers. The factor 1 is the default implementation without a branch multiplier, $N^{-1/2}$ treats the loops as extra depth under Depth-μP, and $N^{-3/2}$ decays faster than $1/N$. Under all three, the best learning rate shifts with N, and the best loss over the learning-rate sweep is higher than under Loop-μP at every loop count above one (Section 4).

Together, our theory and experiments offer a simple recipe for scaling looped Transformers, thus enabling the loop count to join width and depth as a third axis for scaling language models.

2 Related Work

Looped Transformers. Looped Transformers apply the same stack of layers several times, which increases the executed depth without adding parameters. The idea goes back to recurrent depth in Universal Transformers (Dehghani et al., 2019) and cross-layer parameter sharing in ALBERT (Lan et al., 2020). In theory, looping lets fixed weights carry out an iterative computation. Giannou et al. (2023) construct a looped Transformer with fewer than 13 layers that emulates a one-instruction computer and can run general programs. Gatmiry et al. (2024) prove that for in-context linear regression, the best looped linear Transformer implements multi-step preconditioned gradient descent. Saunshi et al. (2025) show that T loops can simulate T steps of chain-of-thought. In practice, looped models match standard Transformers with far fewer parameters when learning function classes in context (Yang et al., 2024b). They also generalize to longer inputs on algorithmic tasks when the loop count grows with the input length (Fan et al., 2025). In language modeling, looping part of a model improves loss at a fixed parameter count (Ng & Wang, 2024). Saunshi et al. (2025) find that looping helps reasoning tasks more than memorization tasks such as closed-book question answering. Zhu et al. (2025) report that accuracy rises with the loop count up to the count used in training, and Geiping et al. (2025) improve reasoning benchmarks by running more loops at test time.

Training stability of deep and looped networks. As depth grows, the contributions of successive layers accumulate, so the forward signal and its updates can explode or vanish. For deep networks with independent layers, one group of fixes changes the architecture. ReZero gates each residual branch with a learnable scalar that starts at zero and removes layer normalization (Bachlechner et al., 2021). DeepNet up-weights the skip connection before each layer normalization and scales down the initialization (Wang et al., 2024). Another group keeps the architecture and changes only the scaling. Fixup rescales the branch weights at initialization according to depth (Zhang et al., 2019). Marion et al. (2025) characterize which residual multipliers keep the signal nontrivial as depth grows. They find that weights varying smoothly across layers require the residual multiplier $1/L$, whose limit is a neural ODE (Chen et al., 2018). Looped models are harder to train, and most existing fixes change the architecture. Geiping et al. (2025) observe that the hidden states in their first large run collapse, and they succeed only after changing the normalization and cutting the learning rate by a factor of ten. Prairie et al. (2026) treat the loop as a dynamical system and constrain the matrix that carries the hidden state from one loop to the next, so that the state cannot explode. Concurrently, DeepLoop (Li et al., 2026) and FPRM (Movahedi et al., 2026) adjust the residual scaling to stabilize looped Transformers, and Chen et al. (2026) damp each residual update by $1/N$ when looping frozen pretrained blocks. None of these derive how hyperparameters should scale with N.

Hyperparameter transfer and parameterization. Tensor Programs derive the maximal-update parameterization, under which hyperparameters tuned on a narrow model transfer to wider ones (Yang & Hu, 2021; Yang et al., 2022), and extend it to adaptive optimizers (Yang & Littwin, 2023). Depth-μP scales each residual branch by $L^{-1/2}$ to maximize feature diversity across layers (Yang et al., 2024a), and Bordelon et al. (2024) and Hayou & Yang (2023) study the same scaling through its training dynamics and its order of limits. CompleteP chooses L^{-1} so that every layer stays non-lazy and reports more reliable depth transfer than with $L^{-1/2}$ (Dey et al., 2025). Later work reports learning-rate transfer under both exponents, with slightly lower loss for $L^{-1/2}$ in two settings (Mlodozeniec et al., 2025; Hägele et al., 2026), and extends the L^{-1} rule to matrix-preconditioned optimizers (Qiu et al., 2025). All of these works scale depth with independent weights, where either exponent in $[1/2, 1]$ transfers the learning rate. We show that the loop axis differs. Because a shared layer adds positively correlated increments at every loop, only the branch multiplier $1/N$ with an N-independent learning rate keeps the forward pass stable and the updates maximal. Loop-μP combines this loop factor with either depth exponent (Table 1).

3 Loop Scaling with One Shared Residual Block

3.1 Setup

We start with a single residual block applied N times with shared weights. Our goal is to determine how the branch multiplier and the learning rate should scale with N. For a Transformer, this corresponds to a single attention or MLP block, and the corresponding recipe is given in Section 3.3.

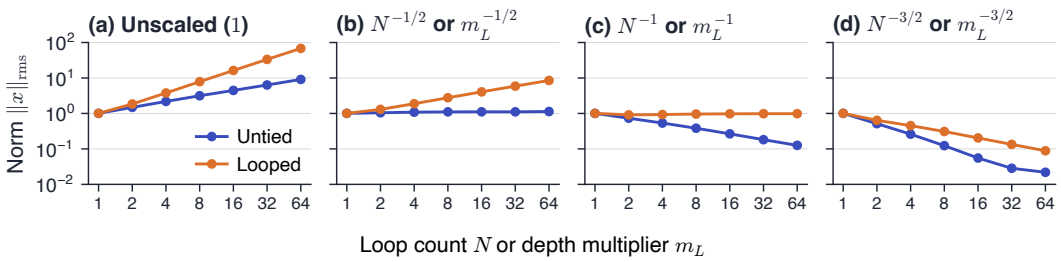


Figure 2: **Residual-stream norm at initialization under different branch multipliers for untied and looped models.** We measure the normalized residual-stream norm $\|\boldsymbol{x}\|_{\text{rms}}$ of Transformers at initialization and plot it (log scale) against the loop count N of a looped model (orange) or the depth multiplier $m_L = L/12$ of an untied model with L layers (blue). The looped model repeats one 12-layer pre-norm Transformer stack N times, and the untied model chains m_L independent copies of that stack. The norm of the untied model stays of order one only under the branch multiplier $m_L^{-1/2}$, while the norm of the looped model stays of order one only under N^{-1}.

Model. Following Depth-μP (Yang et al., 2024a), we use an MLP of width n with a scalar input ζ and a scalar output. The trainable hidden matrix is $\boldsymbol{W}_t \in \mathbb{R}^{n\times n}$, with t indexing training steps; the input and readout matrices $\boldsymbol{U} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times 1}$ stay fixed. Each loop applies a coordinatewise activation ϕ and adds the branch output to the hidden state, scaled by the branch multiplier $c_N = aN^{-\alpha_N}$ for a constant $a > 0$. As in Depth-μP, we subtract the mean of the branch output over its coordinates, $\text{MS}(\boldsymbol{z}) = \boldsymbol{z} - \mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\boldsymbol{z}/n$, where $\mathbf{1}_n$ is the all-ones vector. Set $\boldsymbol{x}_t^0(\zeta) = \boldsymbol{U}\zeta$. The network computes

$$\boldsymbol{h}_t^r(\zeta) = \boldsymbol{W}_t\boldsymbol{x}_t^{r-1}(\zeta), \qquad \boldsymbol{x}_t^r(\zeta) = \boldsymbol{x}_t^{r-1}(\zeta) + c_N\,\text{MS}(\phi(\boldsymbol{h}_t^r(\zeta))), \qquad f_t(\zeta) = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{x}_t^N(\zeta), \quad (1)$$

for $1 \le r \le N$, where $\boldsymbol{x}_t^r$ is the hidden state after r loops and $\boldsymbol{h}_t^r$ is the preactivation in loop r. An untied L-layer network has the same form, except that each layer has its own hidden matrix. The exponent α_N controls how fast the branch multiplier decreases with N. For an untied L-layer network, Depth-μP sets the branch multiplier to $aL^{-1/2}$, corresponding to the unique optimal depth exponents $(\alpha_L, \gamma_L) = (1/2, 1/2)$. We will see that sharing the hidden matrix aligns residual contributions across loops, requiring a loop exponent of 1 rather than $1/2$.

Initialization and limits. We initialize the weights with μP (Yang & Hu, 2021), $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{V}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,n^{-2})$, and $\boldsymbol{W}_{0,ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,n^{-1})$. Under this initialization, each coordinate of the hidden state and preactivation is of order one as $n \to \infty$ at any fixed loop index. We analyze the network in the limit of large width and many loops. As in Depth-μP, we take $n \to \infty$ before $N \to \infty$.

Training. We train only the shared hidden matrix, and keep $\boldsymbol{U}$ and $\boldsymbol{V}$ fixed. At step t, we draw a training input ξ_t and a differentiable loss ℓ_t, and write $\boldsymbol{G}_t = \nabla_{\boldsymbol{W}}\ell_t(f_t(\xi_t))$ for the gradient of $\boldsymbol{W}$. Following Tensor Programs IVb and VI (Yang & Littwin, 2023; Yang et al., 2024a), we rescale $\boldsymbol{G}_t$ by nN^{α_N-1} so that its entries stay of order one (Section 3.2). With a base learning rate $\eta > 0$, an exponent γ_N, and a map Q_t that turns the rescaled gradients into an update direction, the update is

$$\widetilde{\boldsymbol{G}}_t = nN^{\alpha_N-1}\boldsymbol{G}_t, \qquad \boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta n^{-1}N^{-\gamma_N}Q_t(\widetilde{\boldsymbol{G}}_0,\dots,\widetilde{\boldsymbol{G}}_t). \quad (2)$$

For SGD, Q_t returns the last rescaled gradient, whereas for Adam it applies entrywise the bias-corrected ratio $\widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon_{\text{opt}})$ of the first- and second-moment estimates of the rescaled-gradient history. The question of this section is which pair (α_N, γ_N) to use.

3.2 Desiderata for hyperparameter transfer

We require a stable forward pass at initialization, and training updates that keep the hidden-state update bounded while producing an order-one output response. To read off the exponents, we take the identity activation, drop mean subtraction, and study one nondegenerate SGD step. All scales below take $n \to \infty$ before $N \to \infty$, at fixed inputs, and we use the RMS norm $\|\boldsymbol{z}\|_{\text{rms}} = \|\boldsymbol{z}\|/\sqrt{n}$.

Forward pass at initialization. The first requirement keeps the network's representations and output bounded as the loop count grows.

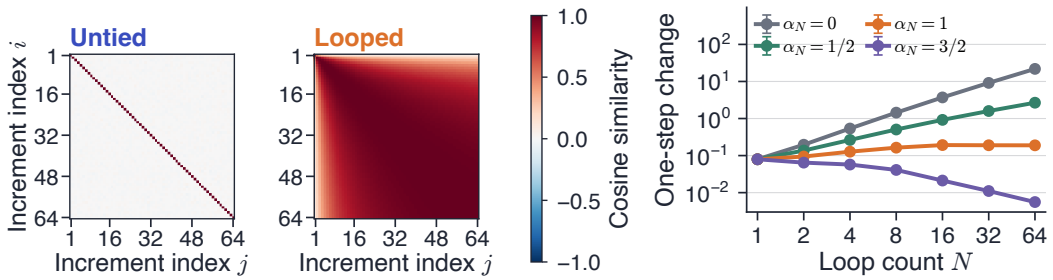


Figure 3: **Left:** we measure in Transformers at initialization the cosine similarity between the residual-stream increments $\boldsymbol{x}^r - \boldsymbol{x}^{r-1}$ of successive 12-layer stacks, without residual scaling. In an untied model of 64 independent stacks the increments are nearly orthogonal, while in a looped model with N=64 they are all positively correlated. **Right:** we measure the RMS change of the final residual stream after one AdamW step at a learning rate shared across N. Setup as in Section 4.1.

Desideratum 1 (Stable initialization)**.** For every fixed bounded input ζ, the initialized network should satisfy that all three of $\|\boldsymbol{x}_0^r(\zeta)\|_{\mathrm{rms}}$, $\|\boldsymbol{h}_0^r(\zeta)\|_{\mathrm{rms}}$, and $|f_0(\zeta)|$ are $\mathcal{O}(1)$, uniformly over $1 \le r \le N$.

Take the input $\zeta = 1$. With a shared matrix, the forward map becomes $\boldsymbol{x}_0^N(1) = (\boldsymbol{I} + c_N \boldsymbol{W}_0)^N \boldsymbol{U}$, and every way of picking m residual terms produces the *same* vector $\boldsymbol{W}_0^m \boldsymbol{U}$. At fixed N, these vectors become orthogonal in normalized inner product as $n \to \infty$, so

$$\boldsymbol{x}_0^N(1) = \sum_{m=0}^{N} \binom{N}{m} c_N^m \boldsymbol{W}_0^m \boldsymbol{U}, \qquad \|\boldsymbol{x}_0^N(1)\|_{\mathrm{rms}}^2 \xrightarrow{\mathbb{P}} \sum_{m=0}^{N} \binom{N}{m}^2 c_N^{2m}.$$

If the layers are instead independent, distinct products remain orthogonal and the squared norm tends to $\sum_{m=0}^{N} \binom{N}{m} c_N^{2m} = (1 + c_N^2)^N$. The linear increment ($m = 1$) thus has size $N c_N$ under shared loops and only $\sqrt{N} c_N$ under untied layers.

With $\beta_N = N^{1-\alpha_N}$, the linear residual contribution has scale $N c_N = a\beta_N$. Thus the initialized residual stream grows with N when $\alpha_N < 1$, stays of order one when $\alpha_N = 1$, and vanishes toward the input $\boldsymbol{U}$ when $\alpha_N > 1$. Desideratum 1 therefore needs $\alpha_N \ge 1$. In our experiments on the pre-norm Transformer (Figures 2 and 3), the same accumulation mechanism is at work: residual increments across loops are more aligned than those across independent 12-layer stacks, producing linear rather than square-root norm growth without a branch multiplier (Section 4.1).

One gradient step. Write $\Delta \boldsymbol{x}_0^r(\zeta) = \boldsymbol{x}_1^r(\zeta) - \boldsymbol{x}_0^r(\zeta)$, and define $\Delta \boldsymbol{h}_0^r$ and Δf_0 likewise. The second requirement asks for bounded updates to the residual stream and order-one output responses.

Desideratum 2 (Stable training and nontrivial response)**.** For bounded inputs and loss derivatives, a single gradient update should keep $\|\Delta \boldsymbol{x}_0^r(\zeta)\|_{\mathrm{rms}}$ and $\|\Delta \boldsymbol{h}_0^r(\zeta)\|_{\mathrm{rms}}$ uniformly $\mathcal{O}(1)$ over $1 \le r \le N$. For a nonzero input and a nonzero loss derivative, it should additionally satisfy $|\Delta f_0(\zeta)| = \Theta(1)$.

We now work in the initialization-stable regime $\alpha_N \ge 1$. For a nonzero input ξ_0 and a nonzero initial loss derivative $\chi_0 = \ell_0'(f_0(\xi_0))$, the shared gradient collects contributions from all loops:

$$\boldsymbol{G}_0 = c_N \sum_{r=1}^{N} \frac{\partial \ell_0}{\partial \boldsymbol{x}_0^r}(\xi_0)\, \boldsymbol{x}_0^{r-1}(\xi_0)^{\mathsf{T}}, \qquad \widetilde{\boldsymbol{G}}_0 = n N^{\alpha_N - 1} \boldsymbol{G}_0.$$

Each backward signal inherits the readout scale $1/n$, and the N aligned terms bring an extra factor $N c_N = a\beta_N$, so the entries of $\boldsymbol{G}_0$ are of order β_N/n. Multiplying by $nN^{\alpha_N - 1}$ then yields $\widetilde{\boldsymbol{G}}_0 = \Theta(1)$. An SGD step therefore changes each entry of $\boldsymbol{W}$ by order s_N/n, where $s_N = N^{-\gamma_N}$.

The first loop receives the raw input $\boldsymbol{U}\zeta$, giving the exact direct response $\Delta \boldsymbol{h}_0^1(\zeta) = \Delta \boldsymbol{W}_0 \boldsymbol{U} \zeta$. The gradient is aligned with $\boldsymbol{U}$, so this response has RMS scale s_N. Boundedness of this first loop therefore requires $\gamma_N \ge 0$. The same $\Delta \boldsymbol{W}_0$ is then reused in every loop and multiplied by c_N each time, so the aggregate response $|\Delta f_0(\zeta)|$ has scale $N c_N s_N = a\rho_N$, where $\rho_N = \beta_N s_N = N^{1-\alpha_N-\gamma_N}$. The first-step output response grows with N if $\alpha_N + \gamma_N < 1$, vanishes if $\alpha_N + \gamma_N > 1$, and stays of order one if and only if $\alpha_N + \gamma_N = 1$. The first-step measurements in Figure 3 illustrate how the loop factor $N^{-\alpha_N}$ controls the response (Section 4.1).

Table 1: **Recipe of Loop-μP.** The left column is the untied recipe of Dey et al. (2025), and the right column shows its looped counterpart. Width, depth, and loop terms are shown in green, blue, and orange. Loop-μP uses $\alpha_L = 1/2$. Rows marked ‡ use the loop-depth theory (Appendix B.4), and rows marked † retain the untied recipe. The base values σ, η, ω, and ϵ_{opt} are tuned on the base model, where both columns coincide.

	Untied, L layers	Looped, L layers $\times$ N loops
Emb. init. var.†	σ^2	σ^2
Emb. LR†	η	η
Pre-LN init. var.†	σ^2	σ^2
Pre-LN LR†	$\eta\, m_L^{\alpha_L - 1}$	$\eta\, m_L^{\alpha_L - 1}$
Hidden init. var.‡	$\sigma^2\, m_n^{-1}$	$\sigma^2\, m_n^{-1}$
Hidden LR‡	$\eta\, m_n^{-1}\, m_L^{\alpha_L - 1}$	$\eta\, m_n^{-1}\, m_L^{\alpha_L - 1}$
Hidden bias LR†	$\eta\, m_L^{\alpha_L - 1}$	$\eta\, m_L^{\alpha_L - 1}$
Hidden WD†	$\omega\, m_n$	$\omega\, m_n$
MHA residual‡	$\boldsymbol{X} + m_L^{-\alpha_L}\, \mathrm{MHA}(\mathrm{LN}(\boldsymbol{X}))$	$\boldsymbol{X} + N^{-1}\, m_L^{-\alpha_L}\, \mathrm{MHA}(\mathrm{LN}(\boldsymbol{X}))$
MLP residual‡	$\boldsymbol{Z} + m_L^{-\alpha_L}\, \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Z}))$	$\boldsymbol{Z} + N^{-1}\, m_L^{-\alpha_L}\, \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Z}))$
Final-LN init. var.†	σ^2	σ^2
Final-LN LR†	η	η
Unemb. init. var.†	σ^2	σ^2
Unemb. LR†	η	η
Unemb. forward†	$\boldsymbol{X}\boldsymbol{W}_{\text{unemb}}^{\mathsf{T}}\, m_n^{-1}$	$\boldsymbol{X}\boldsymbol{W}_{\text{unemb}}^{\mathsf{T}}\, m_n^{-1}$
AdamW ϵ_{opt}, residual blocks‡	$\epsilon_{\text{opt}}\, m_n^{-1}\, m_L^{-\alpha_L}$	$\epsilon_{\text{opt}}\, m_n^{-1}\, m_L^{-\alpha_L}$
AdamW ϵ_{opt}, emb. and unemb.†	$\epsilon_{\text{opt}}\, m_n^{-1}$	$\epsilon_{\text{opt}}\, m_n^{-1}$

Together, Desiderata 1 and 2 give $\alpha_N \geq 1$, $\gamma_N \geq 0$, and $\alpha_N + \gamma_N = 1$, whose only solution is $(\alpha_N, \gamma_N) = (1, 0)$. The branch multiplier is therefore a/N. Substituting into Eq. (2), the raw SGD rate $\eta N^{\alpha_N - 1 - \gamma_N}$ collapses to η; Adam keeps order-one coordinates inside Q_t and uses the external rate $\eta n^{-1} N^{-\gamma_N} = \eta/n$. Both rates are independent of N. With this choice, one training step with sufficiently small $\eta > 0$ produces order-one changes in the hidden states and preactivations at every fixed positive loop fraction: for any $\lambda \in (0, 1]$, both $\|\Delta \boldsymbol{x}_0^{\lfloor \lambda N \rfloor}(\zeta)\|_{\text{rms}}$ and $\|\Delta \boldsymbol{h}_0^{\lfloor \lambda N \rfloor}(\zeta)\|_{\text{rms}}$ are $\Theta(1)$ on a fixed nonzero input. We defer the formal classification of parameterizations and the corresponding feature-learning statements to Appendix A.

3.3 Extension to Looped Transformers

Theory roadmap. We derive $(\alpha_N, \gamma_N) = (1, 0)$ as the unique choice of loop exponents by analyzing one-step SGD training of a single-layer residual network in Section 3.2. In Appendix A, we generalize this result to any finite training horizon by classifying the space of (α_N, γ_N) (Claims 1–3). In Appendix A.3, we then extend the analysis to multi-layer residual branches (e.g., two-layer MLPs) (Claims 2′ and 3′). When the looped object is an untied L-layer residual network (Appendix B), we introduce additional depth exponents (α_L, γ_L) as in Depth-μP, taking $L \to \infty$ before $N \to \infty$. Claims 4–6 then single out the loop-depth parameterization we call Loop-μP: $(\alpha_N, \alpha_L; \gamma_N, \gamma_L) = \left(1, \frac{1}{2}; 0, \frac{1}{2}\right)$. Figures 7 and 8 give the corresponding phase diagrams. We translate this choice into the Transformer recipe in Appendix B.4. The linear-network arguments under SGD are proved in Appendices F–H. Appendix I sketches how the same exponents persist under extra assumptions on nonlinear activations and Adam.

Recipe for looped Transformers. Following Dey et al. (2025), we work with a pre-LN Transformer of L distinct layers, each consisting of one attention branch and one MLP branch, and we reuse the stack for N loops. The resulting $2L$ residual branches are the architectural counterparts of the blocks analyzed in Appendix B. We write $m_n = n/n_{\text{base}}$ and $m_L = L/L_{\text{base}}$ for the width and depth multipliers relative to a base model. The depth exponents (α_L, γ_L) control how the branch multiplier and the learning rate scale with m_L, while (α_N, γ_N) do the same for N. Because m_L is proportional to L, the depth factor $m_L^{-\alpha_L}$ matches the factor $L^{-\alpha_L}$ of Appendix B up to the constant a. As in Dey et al. (2025), we allow $\alpha_L \in [1/2, 1]$, and Loop-μP takes $\alpha_L = 1/2$.

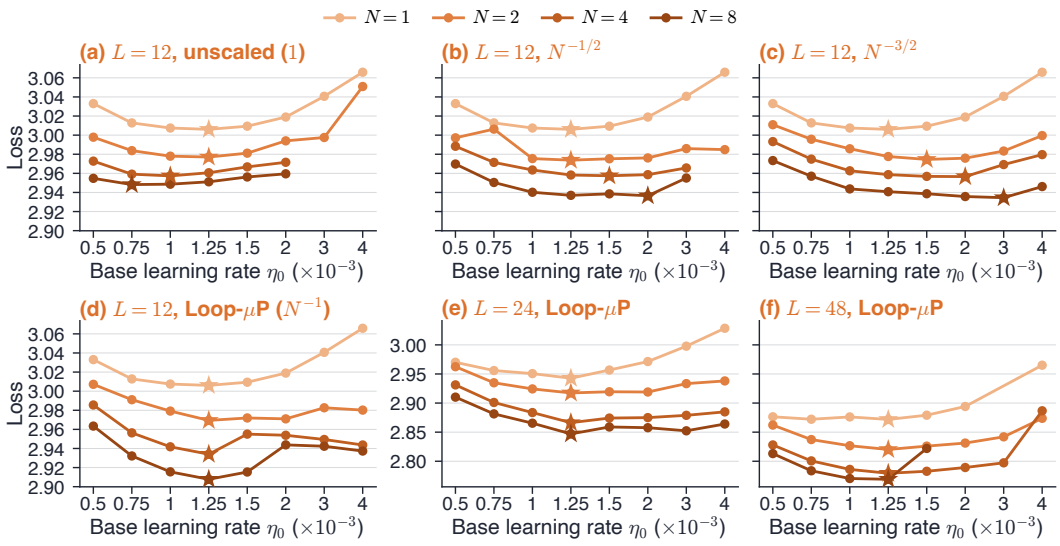


Figure 4: **Only Loop-μP keeps one base learning rate best across loop counts and depths.** Final validation loss against the base learning rate η_0 for loop counts $N \in \{1, 2, 4, 8\}$; stars mark the best learning rate for each N. Panels (a)–(c) use the loop factors 1, $N^{-1/2}$, and $N^{-3/2}$ at L=12, and panels (d)–(f) apply Loop-μP (N^{-1}) at L=12, 24, and 48. The hidden learning rate is $\eta_0 m_L^{-1/2}$ with $m_L = L/12$, so it equals η_0 in panels (a)–(d) and is smaller than η_0 in panels (e) and (f). The best η_0 stays at 1.25×10^{-3} for every N under Loop-μP (panels (d)–(f)), whereas it moves with N under the other loop factors (panels (a)–(c)). Diverged runs (final validation loss > 4) are omitted.

Table 1 compares the untied recipe (left) with its looped version (right). Motivated by Section 3.2, we multiply each shared residual branch by an extra N^{-1} on top of the untied depth factor, which gives the branch multiplier $aN^{-1}m_L^{-\alpha_L}$. Under Loop-μP, the branch multiplier is $aN^{-1}m_L^{-1/2}$ and the Adam learning-rate factor on hidden weights is $m_n^{-1}m_L^{-1/2}$. Treating the N loops as ordinary extra depth would instead scale the branch as $N^{-1/2}m_L^{-1/2}$, and would allow the hidden state to grow with N (Figure 2, panel (b)). Apart from the branch multiplier, the two columns share the same initialization, learning-rate, weight-decay, and AdamW ϵ_{opt} rules, so a model tuned at $N = 1$ can be looped by rescaling only its residual branches.

4 Experiments

We now test whether the Loop-μP recipe in Table 1 transfers hyperparameters across loop counts when training looped Transformer language models. All models are decoder-only Transformers with a Llama-style pre-norm architecture (Vaswani et al., 2017; Touvron et al., 2023). We keep the token embedding and unembedding head outside the loop and repeat a stack of L distinct layers for N loops, and we call each loop a pass. Each layer has two residual branches (attention and MLP), and we scale each branch by the loop factor $N^{-\alpha_N}$. Each branch plays the role of the residual block in Section 3, so the two branches per layer only change the constant a. We first check the predicted residual scaling in the forward pass and the first optimizer step (Section 4.1), then train language models to test learning-rate transfer across loop counts and depths (Section 4.2), and finally check that the mechanism still holds after training (Section 4.3).

4.1 Residual scaling in the forward pass and after one optimizer step

Protocol. We run a controlled experiment that tests the two requirements of Section 3.2, a bounded forward pass at initialization and an order-one change after one optimizer step. We vary the loop count $N \in \{1, 2, 4, 8, 16, 32, 64\}$ and compare four loop factors, 1 (α_N=0), $N^{-1/2}$ (α_N=1/2), N^{-1} (α_N=1), and $N^{-3/2}$ (α_N=3/2). The looped model repeats one 12-layer stack. As a reference, we build untied models of L layers by chaining $m_L = L/12$ independent 12-layer stacks, and we scale their branches by the depth factor $m_L^{-\alpha_L}$. We write $\boldsymbol{x}^r$ for the residual stream after the r-th 12-layer

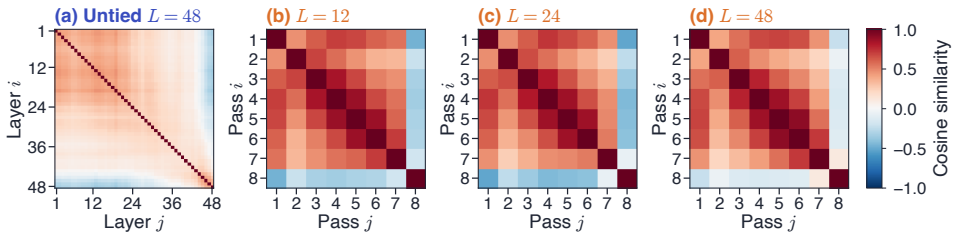


Figure 5: **After training, the pass increments of looped models stay aligned, while the per-layer increments of an untied model align mainly with nearby layers.** Cosine similarity between residual-stream increments after full training (20,000 steps), for the 48 layers of an untied L=48 model (a) and for the 8 passes $\boldsymbol{x}^r - \boldsymbol{x}^{r-1}$ of looped models with N=8 and $L \in \{12, 24, 48\}$ (b–d). In the looped models, every pair of passes that excludes the final pass is positively correlated, whereas in the untied model the similarity fades with layer distance.

stack, which in a looped model is the r-th pass. All runs use batch size 1, sequence length 128, and fixed random token inputs, and we average over 10 seeds. For the forward pass, we measure at initialization the RMS norm $\|\boldsymbol{x}^N\|_{\mathrm{rms}}$ after the last pass and the pairwise cosine similarities between the increments $\boldsymbol{x}^r - \boldsymbol{x}^{r-1}$. For the first step, we take one AdamW step with a base learning rate shared across all N and measure the RMS change of $\boldsymbol{x}^N$.

Forward pass. *A looped model needs the loop factor N^{-1} to keep the forward pass bounded, whereas an untied model needs only the depth factor $m_L^{-1/2}$.* Without residual scaling, the looped norm grows linearly in N, while the untied norm grows only as $\sqrt{m_L}$ (Figure 2, panel (a)). The untied increments are nearly orthogonal and add in quadrature, whereas the shared weights make the looped increments positively correlated, so they add along one direction (Figure 3, left), as Section 3.2 predicts. Hence $N^{-1/2}$ still lets the looped norm grow with N (panel (b)), and only N^{-1} keeps it approximately invariant across N (panel (c)). The factor $N^{-3/2}$ also bounds the norm but drives the looped model toward the identity map (panel (d)), so the forward pass rules out $\alpha_N < 1$ (Claim 1) without separating α_N=1 from α_N=3/2.

One optimizer step. *With a learning rate shared across N, only N^{-1} gives a one-step change that neither grows nor vanishes with N.* The one-step change grows with N under α_N=0 and α_N=1/2, shrinks under α_N=3/2, and stays within a small range under α_N=1 (Figure 3, right), and the first ten steps show the same pattern (Appendix C.1). As in the forward pass, weight sharing explains this pattern. The same weight update enters all N passes and its effects add along one direction, while AdamW sets the size of the update by the learning rate alone, so the change scales as $\eta N^{1-\alpha_N}$, which is of order one only at α_N=1 (Claim 2). Together with the forward pass, this leaves N^{-1} as the only loop factor under which a learning rate independent of N changes a bounded residual stream by order one, as learning-rate transfer across loop counts requires.

4.2 Learning-rate transfer across loop counts and depths

Protocol. We train looped Transformers with $L \in \{12, 24, 48\}$ distinct layers on FineWeb-Edu (Penedo et al., 2024) for 10B tokens with loop counts $N \in \{1, 2, 4, 8\}$, and for each (L, N) we sweep learning rates around a base value tuned at L=12, N=1. At L=12, we compare four loop factors, 1, $N^{-1/2}$, N^{-1} (Loop-μP), and $N^{-3/2}$. The factor 1 leaves the branches unscaled, and the other three decay more slowly than, as fast as, and faster than the theory prescribes. At L=24 and L=48, we use Loop-μP and, following the prediction $\eta \propto L^{-1/2}$ (Table 1), scale the learning rate by $m_L^{-1/2} = (L/12)^{-1/2}$ (Appendix E.2). The optimizer, schedule, and full model configuration are in Table 2, and we report the validation loss on the FineWeb-Edu held-out split at the end of training.

Transfer across loop counts. *Only N^{-1} keeps the best learning rate fixed across N, and it also reaches the lowest loss.* Under N^{-1}, the optimal learning rate stays at 1.25×10^{-3} for every N at L=12 (Figure 4, panel (d)), and it also stays fixed across N at L=24 and L=48 (panels (e) and (f)). Without a loop factor, the optimum falls from 1.25×10^{-3} at $N \leq 2$ to 7.5×10^{-4} at N=8, and learning rates of 3×10^{-3} and above diverge at $N \geq 4$ (panel (a)), consistent with a one-step change that grows with N (Section 4.1). Under $N^{-3/2}$, the optimum rises from 1.25×10^{-3} at N=1 to

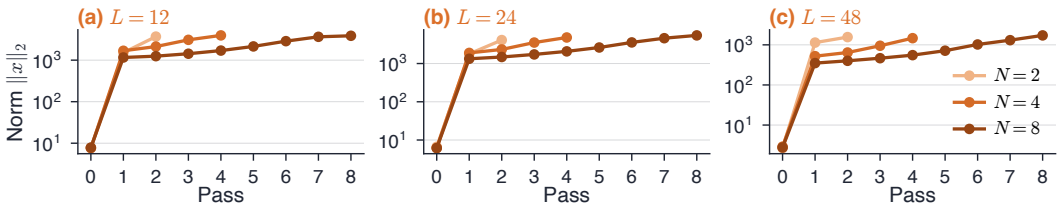


Figure 6: **Under Loop-μP the trained residual stream ends at a similar norm for every loop count.** Each line traces the L2 norm of the residual stream from pass 0 (after the embedding) through the last pass, for $N \in \{2, 4, 8\}$ after full training.

3×10^{-3} at N=8 (panel (c)), so each loop count needs its own sweep, and under $N^{-1/2}$ it also moves at $N \geq 4$ (panel (b)). The theory predicts the rise under $N^{-3/2}$, because at a fixed learning rate the one-step response scales as $\rho_N = N^{1-\alpha_N} = N^{-1/2}$ (Section 3.2), so an order-one response needs a learning rate growing as $N^{1/2}$, which reaches about 3.5×10^{-3} at N=8. Beyond transfer, N^{-1} reaches a lower best loss than all three alternatives at every $N > 1$, and its best loss falls the most as N grows (panels (a)–(d)). This matches the forward pass, where 1 and $N^{-1/2}$ leave the residual stream unbounded and $N^{-3/2}$ drives the looped model toward the identity map.

Transfer across depths. *With the learning rate scaled by $m_L^{-1/2}$, the same base learning rate is also optimal at every depth.* After the depth correction, the optimal base learning rate is the same at L=12, 24, and 48 (Figure 4, panels (d)–(f)). Together with the fixed optimum across N, one base learning rate is optimal for every (L, N) we train, which shows Loop-μP's effectiveness.

4.3 Residual stream beyond initialization

Protocol. We check whether the two predictions that we verified at initialization in Section 4.1 still hold after full training (20,000 steps) of the $L \in \{12, 24, 48\}$ models from Section 4.2. For N=8, we compute the cosine similarity between the increments of every pair of passes, and for comparison we compute the same matrix between the per-layer increments of the trained untied L=48 model (Figure 5). For $N \in \{2, 4, 8\}$, we trace the norm of the residual stream after every pass, from the embedding output (r=0) through the last pass (Figure 6).

Increment alignment. *After training, the increments of different passes stay aligned.* We find that the cosine similarities between pass increments are lower than at initialization (Figure 3) but remain positive for every pair of passes that excludes the final pass, at all three depths (Figure 5, panels (b)–(d)) and also at N=4 (Appendix C.2). Only the increment of the final pass is weakly or negatively correlated with the others. In contrast, we find that the per-layer increments of the trained untied model have low similarity overall and align only with nearby layers (panel (a)), as in the untied models with L=12 and L=24 (Appendix C.3).

Final norm. *Under Loop-μP, the final residual norm is consistent across loop counts.* We find that the per-pass trajectories differ across N, but within each depth the norm after the last pass varies by at most a factor of 1.33, as it stays approximately invariant at initialization (Figure 6).

5 Conclusion

We study how to train looped Transformers stably as the number of loops grows. By analyzing a residual block looped N times in the infinite-width limit, we find that looping differs from adding depth, because the loops share weights and their updates to the residual stream therefore add along the same direction rather than in independent directions. This property leads to Loop-μP, which multiplies each residual branch by $1/N$ on top of an existing depth parameterization and keeps the learning rate independent of N. In our language-modeling experiments, Loop-μP trains looped Transformers stably, transfers the learning rate across loop counts up to 8 and executed depths up to 384 layers, and reaches a lower loss than loop factors that decay more slowly or faster than $1/N$. We hope that Loop-μP offers the looped-model community a simple recipe for training better looped models at a lower tuning cost. We discuss the limitations of this work in Appendix D.

AI USE STATEMENT

We used generative AI tools to edit the paper for readability and to write and edit parts of the plotting and analysis code. We also used these tools to help verify the mathematical proofs. The authors reviewed all AI-edited text, tested the AI-assisted code, and checked every AI-assisted figure against the underlying experimental data. We take responsibility for the final content of this work, including text, code, and figures produced with the aid of generative AI.

ETHICS STATEMENT

This work studies how to parameterize looped Transformers and involves no human subjects or personal data. We train only on the publicly released FineWeb-Edu corpus (Penedo et al., 2024) and release no new datasets or model weights, so we do not foresee ethical concerns beyond those common to language-model research.

REPRODUCIBILITY STATEMENT

Table 1 gives the Loop-μP recipe, and Appendix E lists all model configurations and hyperparameters. We provide our code in the supplementary material.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Gary Cottrell, and Julian J. McAuley. ReZero is all you need: fast convergence at large depth. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, Proceedings of Machine Learning Research, pp. 1352–1361. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KZJehvRKGD.

Lizhang Chen, Jonathan Li, Chen Liang, Ni Lao, and Qiang Liu. Training-free looped transformers. *arXiv preprint arXiv:2605.23872*, 2026. doi: 10.48550/arXiv.2605.23872.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6572–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: CompleteP enables compute-efficient deep transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theory. *Annales de l'I.H.P. Probabilités et statistiques*, 23(S2):397–423, 1987.

Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. Looped transformers for length generalization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=2edigk8yoU.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 15130–15152. PMLR / OpenReview.net, 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. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. doi: 10.48550/arXiv.2502.05171.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning, ICML 2023*, 2023.

Alexander Hägele, Alejandro Hernández-Cano, Atli Kosson, and Martin Jaggi. Improving neural network training by decoupling the magnitude and direction of weight vectors. *arXiv preprint arXiv:2606.25971*, 2026. doi: 10.48550/arXiv.2606.25971.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 12700–12723. PMLR, 2023. URL https://proceedings.mlr.press/v202/hayou23a.html.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Shuzhen Li, Yifan Zhang, Jiacheng Guo, Quanquan Gu, and Mengdi Wang. DeepLoop: Depth scaling for looped transformers. *arXiv preprint arXiv:2607.13491*, 2026. doi: 10.48550/arXiv.2607.13491.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pierre Marion, Adeline Fermanian, Gérard Biau, and Jean-Philippe Vert. Scaling resnets in the large-depth regime. *J. Mach. Learn. Res.*, 26:56:1–56:48, 2025. URL https://jmlr.org/papers/v26/22-0664.html.

Bruno Mlodozeniec, Pierre Ablin, Louis Béthune, Dan Busbridge, Michal Klein, Jason Ramapuram, and Marco Cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. 2025. URL https://arxiv.org/abs/2512.22382.

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. *arXiv preprint arXiv:2606.18206*, 2026. doi: 10.48550/arXiv.2606.18206.

Kei-Sing Ng and Qingchen Wang. Loop neural networks for parameter sharing. 2024. URL https://arxiv.org/abs/2409.14199.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models. *arXiv preprint arXiv:2604.12946*, 2026. doi: 10.48550/arXiv.2604.12946.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pp. 157–163. Association for Computational Linguistics, 2017. doi: 10.18653/V1/E17-2025. URL https://doi.org/10.18653/v1/e17-2025.

Shikai Qiu, Zixi Chen, Hoang Phan, Qi Lei, and Andrew Gordon Wilson. Hyperparameter transfer enables consistent gains of matrix-preconditioned optimizers across scales. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. doi: 10.48550/arXiv.2512.05620.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=din0lGfZFd.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling transformers to 1,000 layers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(10):6761–6774, 2024. doi: 10.1109/TPAMI.2024.3386927. URL https://doi.org/10.1109/TPAMI.2024.3386927.

Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *CoRR*, abs/2410.05192, 2024. doi: 10.48550/ARXIV.2410.05192. URL https://doi.org/10.48550/arXiv.2410.05192.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 10524–10533. PMLR, 2020. URL http://proceedings.mlr.press/v119/xiong20b.html.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, volume 139, pp. 11727–11737, 2021.

Greg Yang and Etai Littwin. Tensor programs IVb: Adaptive optimization in the infinite-width limit, 2023. URL https://arxiv.org/abs/2308.01814.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. 2022. doi: 10.48550/ARXIV.2203.03466. URL https://arxiv.org/abs/2203.03466.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite-depth neural networks. In *International Conference on Learning Representations*, 2024a.

Liu Yang, Kangwook Lee, Robert D. Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=HHbRxoDTxE.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1gsz30cKX.

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. doi: 10.48550/ARXIV.2510.25741. URL https://arxiv.org/abs/2510.25741.

APPENDIX CONTENTS

A Formal Classification of Loopwise Parameterizations **16**
A.1 Definitions . . . 16
A.2 Classification at block depth 1 . . . 17
A.3 Classification at fixed block depth . . . 17

B Loop-μP for Stacks of Distinct Layers **19**
B.1 Setup . . . 19
B.2 Desiderata for loop-depth hyperparameter transfer . . . 19
B.3 Classification of loop-depth parameterizations . . . 21
B.4 From Loop-μP to the Transformer recipe . . . 22

C Additional Experimental Results **22**
C.1 One-step response over the first ten steps . . . 23
C.2 Post-training cosine similarity at a smaller loop count . . . 23
C.3 Post-training layer increments of shallower untied models . . . 23

D Limitations **24**

E Experimental Setup and Implementation Details **24**
E.1 Language modeling experiment configuration . . . 24
E.2 Loop-μP parameterization used in the experiments . . . 24
E.3 Hyperparameter values and sweep configuration . . . 25

F Notation and Mathematical Tools **25**

G Proofs for the Infinite-Loop Limit in the Linear Case **27**
G.1 SGD training dynamics . . . 28
G.2 Classification of loopwise parameterizations at fixed block depth . . . 35
G.3 Sharper nontriviality classification at block depth 1 . . . 38
G.4 A block-depth-2 counterexample outside training stability . . . 40

H Proofs for the Infinite Loop-Depth Limit in the Linear Case **41**
H.1 SGD training dynamics . . . 43
H.2 Classification of loop-depth parameterizations . . . 47

I Heuristics for the Proofs in the General Case **51**
I.1 Setup and assumptions . . . 51
I.2 Basic properties . . . 54
I.3 Justification of initialization stability . . . 55
I.4 Justification of training stability and nontriviality . . . 56

I.5 Justification of feature properties and extensions 59

Guide to the appendix. Appendix A provides a rigorous statement of the derivations in Section 3. Appendices A.3 and B extend the theory of Section 3 to the models summarized in Section 3.3. Appendix A.3 allows each residual branch to contain several matrices, and Appendix B stacks distinct layers and derives Loop-μP. Appendix C reports additional experimental results for Section 4, Appendix D discusses the limitations of this work, and Appendix E describes the experimental setup. Appendices F–I contain the mathematical tools, the proofs of the Claims for linear networks trained with SGD, and a justification of these Claims for nonlinear networks and Adam.

A Formal Classification of Loopwise Parameterizations

We now formalize the scaling analysis of Section 3. We first introduce the definitions shared by a single-matrix residual branch and by residual branches of fixed finite block depth, and then state the corresponding classification claims. The strict linear-SGD proofs are given in Appendix G; extensions to nonlinear networks and entrywise optimizers are discussed in Appendix I.

A.1 Definitions

We use the following definitions for the model with a single matrix in its residual branch, introduced in Section 3.1, and for the model with a residual branch of fixed finite block depth, introduced in Appendix A.3. In the latter setting, preactivation and feature-learning requirements are imposed at every internal position of the residual branch.

We formulate the definitions for finite input batches. At step t, we train on a batch $\Xi_t = (\xi_{t,b})_{b=1}^{B}$ of scalar inputs; the singleton case $B = 1$ recovers the input ξ_t used in the setup. For evaluation, we fix a finite batch $\boldsymbol{\zeta}$ of scalar inputs, which need not coincide with a training batch, and stack the corresponding hidden states, preactivations, and outputs columnwise into $\boldsymbol{X}_t^r(\boldsymbol{\zeta})$, $\boldsymbol{H}_t^r(\boldsymbol{\zeta})$, and $\boldsymbol{f}_t(\boldsymbol{\zeta})$. For a batch matrix, we use the RMS norm $\|\boldsymbol{X}\|_{\mathrm{rms}} = n^{-1/2}\|\boldsymbol{X}\|_{\mathrm{F}}$, which reduces to the vector RMS norm for a singleton batch. For the fixed-size output batch $\boldsymbol{f}_t(\boldsymbol{\zeta})$, asymptotic statements are understood coordinatewise. We write $\Delta F_t(\boldsymbol{\zeta}) = F_{t+1}(\boldsymbol{\zeta}) - F_t(\boldsymbol{\zeta})$ for the one-step change of any of these quantities. Unless stated otherwise, all asymptotic notation and convergence below refer to the ordered limit $n \to \infty$ followed by $N \to \infty$; the conventions are summarized in Appendix F.

We encode each finite training run by the following data.

Definition 1 (Training routine)**.** A training routine $\mathcal{R} = (T, \eta, \{Q_t, \Xi_t, \ell_t\}_{t<T})$ consists of a finite horizon T, a base learning rate $\eta > 0$, and, at each step $t < T$, an entrywise update map Q_t, a training batch Ξ_t, and a differentiable loss ℓ_t including the labels. The routine does not depend on n or N.

We call a parameterization stable when its initialization and every one-step update remain bounded.

Definition 2 (Stability)**.** We say that a parameterization is

(i) *stable at initialization* if, for every evaluation batch $\boldsymbol{\zeta}$, $\boldsymbol{H}_0^r(\boldsymbol{\zeta}), \boldsymbol{X}_0^r(\boldsymbol{\zeta}) = \mathcal{O}(1)$ for $1 \le r \le N$ and $\boldsymbol{f}_0(\boldsymbol{\zeta}) = \mathcal{O}(1)$;

(ii) *stable during training* if, for every routine $\mathcal{R}$, step $0 \le t < T$, evaluation batch $\boldsymbol{\zeta}$, and $1 \le r \le N$, $\Delta\boldsymbol{H}_t^r(\boldsymbol{\zeta}), \Delta\boldsymbol{X}_t^r(\boldsymbol{\zeta}) = \mathcal{O}(1)$ and $\Delta\boldsymbol{f}_t(\boldsymbol{\zeta}) = \mathcal{O}(1)$.

We call it *stable* when both conditions hold.

Nontriviality asks whether at least one training routine changes the output in the limit.

Definition 3 (Nontriviality)**.** We call a parameterization *trivial* if $\boldsymbol{f}_t(\boldsymbol{\zeta}) - \boldsymbol{f}_0(\boldsymbol{\zeta}) \to 0$ for every routine $\mathcal{R}$, step $1 \le t \le T$, and evaluation batch $\boldsymbol{\zeta}$. It is *nontrivial* otherwise.

We use three further properties to describe the features. Faithfulness requires the preactivations, which are the inputs to ϕ, to remain of order one. To exclude the degenerate zero-input case, we call an evaluation batch nondegenerate when all of its inputs are nonzero.

Definition 4 (Faithfulness)**.** A parameterization is *faithful* on a nondegenerate evaluation batch $\boldsymbol{\zeta}$ if $\boldsymbol{H}_t^r(\boldsymbol{\zeta}) = \Theta(1)$ for every routine $\mathcal{R}$, step $0 \le t \le T$, and loop $1 \le r \le N$, uniformly over r.

Feature learning asks that one training step change the features by order one at every fixed positive loop fraction. We represent the loop position $\lfloor \lambda N \rfloor$ by its fraction $\lambda \in (0, 1]$.

Definition 5 (Feature learning)**.** A parameterization induces *feature learning* if, on a nondegenerate evaluation batch $\boldsymbol{\zeta}$, there exist a routine $\mathcal{R}$ and a step $0 \le t < T$ such that $\Delta \boldsymbol{X}_t^{\lfloor \lambda N \rfloor}(\boldsymbol{\zeta}) = \Theta(1)$ and $\Delta \boldsymbol{H}_t^{\lfloor \lambda N \rfloor}(\boldsymbol{\zeta}) = \Theta(1)$ for every $\lambda \in (0, 1]$.

Unlike in the Depth-μP analysis (Yang et al., 2024a), the critical point for a single-matrix residual branch is already determined by stability and nontriviality. The claims below give $(\alpha_N, \gamma_N) = (1, 0)$ in Appendix A.2, without using feature diversity. We introduce the exponent to describe the loop trajectory at this point. In the loop-depth setting, the corresponding depth-diversity requirement becomes an additional selection criterion; see Appendices B.2 and B.3. Feature diversity measures how rapidly the hidden states vary between nearby loop fractions. Following Depth-μP, we quantify this variation through the scaling of the increment between λ and $\lambda + \varepsilon$ as $\varepsilon \downarrow 0$.

Definition 6 (Feature-diversity exponent)**.** On a nondegenerate evaluation batch $\boldsymbol{\zeta}$, let $\mathcal{K}$ be the set of all $\kappa_N \ge 0$ such that

$$\left\| \boldsymbol{X}_t^{\lfloor (\lambda+\varepsilon)N \rfloor}(\boldsymbol{\zeta}) - \boldsymbol{X}_t^{\lfloor \lambda N \rfloor}(\boldsymbol{\zeta}) \right\|_{\mathrm{rms}} = \Omega(\varepsilon^{1-\kappa_N})$$

for every routine $\mathcal{R}$, step $0 \le t \le T$, $\lambda \in [0, 1)$, and sufficiently small $0 < \varepsilon \le 1 - \lambda$, where the limit is taken first and then $\varepsilon \downarrow 0$. When $\mathcal{K} \neq \varnothing$, we define the feature-diversity exponent to be $\sup \mathcal{K}$. If $\mathcal{K} = \varnothing$, for example because the hidden state does not change across loops, we call the diversity degenerate and leave the exponent undefined.

When matched by a corresponding upper bound, $\kappa_N = 0$ yields the smooth-path scaling $\Theta(\varepsilon)$, whereas larger exponents correspond to stronger local variation.

A.2 Classification at block depth 1

We now formalize the scaling analysis in Section 3.2 for a residual branch containing a single matrix. We use the model of Section 3.1 and the definitions in Appendix A.1. The first two claims characterize initialization stability, training stability, and nontriviality separately.

Claim 1. A parameterization is stable at initialization if and only if $\alpha_N \ge 1$.

Claim 2. Consider a parameterization that is stable at initialization. Training stability and nontriviality are characterized separately as follows:

(i) It is stable during training if and only if $\gamma_N \ge 0$ and $\alpha_N + \gamma_N \ge 1$.
(ii) It is nontrivial if and only if $\alpha_N + \gamma_N \le 1$.

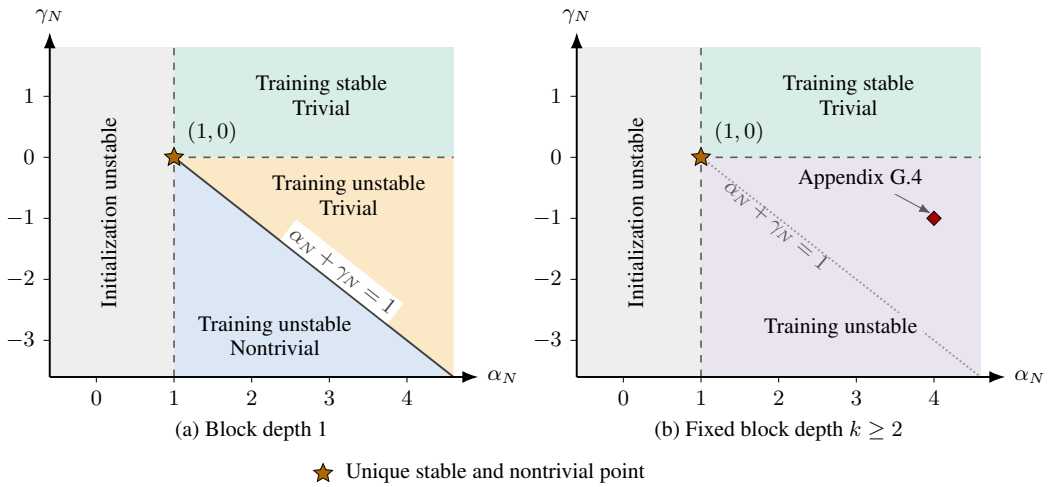
In particular, when $\alpha_N + \gamma_N > 1$, no training routine changes the output in the limit.

Combining the two conditions, a parameterization is stable and nontrivial only at $(\alpha_N, \gamma_N) = (1, 0)$. Figure 7(a) shows the corresponding phase classification. The final claim describes the feature properties at this point.

Claim 3. At the unique stable and nontrivial point $(\alpha_N, \gamma_N) = (1, 0)$, the parameterization is faithful for every training routine. For a sufficiently small $\eta > 0$, a first step on a nonzero training input with a nonzero loss derivative induces feature learning in both the hidden states and the preactivations. Its feature-diversity exponent is $\kappa_N = 0$.

A.3 Classification at fixed block depth

We have so far considered a looped network whose residual branch contains a single matrix (Section 3). We now consider a residual branch of fixed finite block depth k, with k distinct matrices shared across the same N loops. This is the setting relevant to the multi-layer branches in a Transformer (Section 3.3). We first specify the model and update rule, and then explain how the classification changes when the residual branch has depth k.



(a) Block depth 1 (b) Fixed block depth $k \geq 2$

Figure 7: Loopwise classification from Claims 1 and 2 and Claim 2′. Both panels use the same axes and scales. Panel (a) has a single matrix in the residual branch, while panel (b) has a residual branch of fixed block depth $k \geq 2$. In both panels, initialization is stable exactly when $\alpha_N \geq 1$; within this region, training is stable exactly when $\gamma_N \geq 0$. Thus $(1, 0)$ is the only point that is both stable and nontrivial in either setting. In panel (a), the line $\alpha_N + \gamma_N = 1$ is the boundary of the nontrivial region. In panel (b), the dotted line is shown only for comparison with panel (a). The diamond marks the counterexample in Appendix G.4. For residual branches of block depth $k \geq 2$, the nontriviality equivalence is restricted to the training-stable region, so the shaded training-unstable region is not fully classified by that criterion.

Model. Each residual branch contains k matrices, and the same k matrices are used in every loop. We set $\boldsymbol{x}_t^0(\zeta) = \boldsymbol{U}\zeta$ and $\boldsymbol{z}_t^{r,0}(\zeta) = \boldsymbol{x}_t^{r-1}(\zeta)$. With activations $\phi_1, \ldots, \phi_k$, the preactivations $\boldsymbol{h}_t^{r,j}(\zeta)$, internal features $\boldsymbol{z}_t^{r,j}(\zeta)$, and hidden states $\boldsymbol{x}_t^r(\zeta)$ satisfy

$$\begin{aligned} \boldsymbol{h}_t^{r,j} &= \boldsymbol{W}_t^j \boldsymbol{z}_t^{r,j-1}, \\ \boldsymbol{z}_t^{r,j} &= \mathrm{MS}(\phi_j(\boldsymbol{h}_t^{r,j})), \\ \boldsymbol{x}_t^r &= \boldsymbol{x}_t^{r-1} + aN^{-\alpha_N} \boldsymbol{z}_t^{r,k}, \end{aligned} \tag{A1}$$

where $1 \leq r \leq N$ and $1 \leq j \leq k$. The output is $f_t(\zeta) = \boldsymbol{V}^\mathsf{T} \boldsymbol{x}_t^N(\zeta)$. The k matrices are initialized independently with the same Gaussian law as $\boldsymbol{W}_0$ and independently of $\boldsymbol{U}$ and $\boldsymbol{V}$. Each matrix is trained with its gradient $\boldsymbol{G}_{j,t} = \nabla_{\boldsymbol{W}^j} \ell_t(f_t(\xi_t))$, which sums over all N loops, rescaled as in Eq. (2) to $\widetilde{\boldsymbol{G}}_{j,t} = nN^{\alpha_N - 1}\boldsymbol{G}_{j,t}$. All matrices are updated at the same time by

$$\boldsymbol{W}_{t+1}^j = \boldsymbol{W}_t^j - \eta n^{-1} N^{-\gamma_N} Q_t(\widetilde{\boldsymbol{G}}_{j,0}, \ldots, \widetilde{\boldsymbol{G}}_{j,t}). \tag{A2}$$

We apply the definitions in Appendix A.1 with one additional requirement: when the residual branch has depth k, faithfulness and feature learning must hold at every internal position $j = 1, \ldots, k$, as well as for the residual state.

Because k is fixed, the additional matrix products change constants but introduce no new powers of N. Initialization stability and the critical loop exponents therefore remain governed by the same (α_N, γ_N) scales. The nontriviality statement requires a qualification: for residual branches of depth $k \geq 2$, we state nontriviality within the training-stable region, unlike in the single-matrix case covered by Claim 2.

Claim 2′. Consider a parameterization that is stable at initialization. Training stability is characterized first, and nontriviality is then characterized within that stable-training region:

(i) It is stable during training if and only if $\gamma_N \geq 0$.
(ii) If it is also stable during training, it is nontrivial if and only if $\alpha_N + \gamma_N \leq 1$.

Combining these three conditions gives $\alpha_N \geq 1$, $\gamma_N \geq 0$, and $\alpha_N + \gamma_N \leq 1$. Thus the only point that is both stable and nontrivial is $(\alpha_N, \gamma_N) = (1, 0)$. Figure 7(b) shows the corresponding phase classification.

Claim 3′. At the unique stable and nontrivial point $(\alpha_N, \gamma_N) = (1, 0)$, the parameterization is faithful for every training routine at every internal position, and the residual hidden states remain of order one. For a nonzero training input and a nonzero loss derivative, one sufficiently small first step changes the hidden state and every internal preactivation by order one at every positive loop fraction. Its feature-diversity exponent is $\kappa_N = 0$.

Remark 1. The qualification in Claim 2′ (ii) is essential. The counterexample in Appendix G.4 gives a $k = 2$ model that is stable at initialization, has $\gamma_N < 0$ and $\alpha_N + \gamma_N > 1$, yet changes its output by a nonzero amount after two SGD steps. Thus, unlike the $k = 1$ case, nontriviality outside the training-stable region cannot be classified solely by the aggregate response exponent.

B LOOP-μP FOR STACKS OF DISTINCT LAYERS

The recipe in Section 3.3 states Loop-μP for a looped Transformer, which repeats a stack of L distinct layers over N loops, and this appendix derives it. We extend the single-block model of Section 3 to such a stack. The setup, intuitive analysis, classification, and recipe derivation are in Appendices B.1–B.4, respectively. The linear-SGD proof and nonlinear/Adam discussion are given in Appendices H and I.5, respectively.

B.1 SETUP

We now replace the single shared block by an ordered stack of L distinct residual blocks and reuse the whole stack over N loops. Our goal is to determine how the branch multiplier and the learning rate should scale with both L and N. We call one pass through the L blocks a *cycle*.

Model. As in Section 3.1, we use an MLP of width n with a scalar input ζ and a scalar output. The trainable hidden matrices are $\boldsymbol{W}_t^l \in \mathbb{R}^{n\times n}$ for $1 \le l \le L$; the input and readout matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ stay fixed. Each block applies a coordinatewise activation ϕ_l and adds a mean-subtracted branch output to the hidden state, scaled by the branch multiplier $c = aN^{-\alpha_N}L^{-\alpha_L}$ for a constant $a > 0$. Set $\boldsymbol{x}_t^{1,0}(\zeta) = \boldsymbol{U}\zeta$ and $\boldsymbol{x}_t^{r,0}(\zeta) = \boldsymbol{x}_t^{r-1,L}(\zeta)$ for $r \ge 2$. The network computes

$$\boldsymbol{h}_t^{r,l}(\zeta) = \boldsymbol{W}_t^l \boldsymbol{x}_t^{r,l-1}(\zeta), \qquad \boldsymbol{x}_t^{r,l}(\zeta) = \boldsymbol{x}_t^{r,l-1}(\zeta) + c\ \mathrm{MS}\big(\phi_l(\boldsymbol{h}_t^{r,l}(\zeta))\big), \qquad f_t(\zeta) = \boldsymbol{V}^\mathsf{T}\boldsymbol{x}_t^{N,L}(\zeta), \tag{B1}$$

for $1 \le r \le N$ and $1 \le l \le L$. The exponents α_N and α_L control how fast the branch multiplier decreases with the number of loops and with the number of blocks in a cycle.

Initialization and limits. We initialize every hidden matrix with μP as in Section 3.1, independently across l. We analyze the network in the limit of large width, many blocks, and many loops, taking $n \to \infty$ before $L \to \infty$ before $N \to \infty$. Under this initialization, each coordinate of the hidden state and preactivation is of order one as $n \to \infty$ at any fixed pair (r, l).

Training. We train only the hidden matrices and keep $\boldsymbol{U}$ and $\boldsymbol{V}$ fixed. At step t, we draw a training input ξ_t and a differentiable loss ℓ_t, and write $\boldsymbol{G}_t^l = \nabla_{\boldsymbol{W}^l}\ell_t(f_t(\xi_t))$ for the gradient of block l through all N loops. Following Tensor Programs IVb and VI (Yang & Littwin, 2023; Yang et al., 2024a) and Eq. (2), we rescale $\boldsymbol{G}_t^l$ by $nN^{\alpha_N-1}L^{\alpha_L}$ so that its entries stay of order one. With a base learning rate $\eta > 0$, exponents γ_N and γ_L, and a map Q_t that turns the rescaled gradients into an update direction, the update is

$$\widetilde{\boldsymbol{G}}_t^l = nN^{\alpha_N-1}L^{\alpha_L}\boldsymbol{G}_t^l, \qquad \boldsymbol{W}_{t+1}^l = \boldsymbol{W}_t^l - \eta n^{-1}N^{-\gamma_N}L^{-\gamma_L}Q_t(\widetilde{\boldsymbol{G}}_0^l, \dots, \widetilde{\boldsymbol{G}}_t^l). \tag{B2}$$

The exponents γ_N and γ_L control how the learning rate depends on N and L after this rescaling. The question of this appendix is which tuple $(\alpha_N, \alpha_L; \gamma_N, \gamma_L)$ to use.

B.2 DESIDERATA FOR LOOP-DEPTH HYPERPARAMETER TRANSFER

We extend Desiderata 1 and 2 of Section 3.2 from one shared block to a stack of L independent blocks reused over N loops. As there, we take the identity activation, drop mean subtraction, and study one nondegenerate SGD step. All limits below are taken in the order $n \to \infty$, then $L \to \infty$, then $N \to \infty$, with inputs held fixed.

Forward pass at initialization. The first requirement keeps the network's representations and output bounded as L and N grow, which we read off from Desideratum 1 applied to every pair (r, l).

Take the input $\zeta = 1$. Within one cycle the L hidden matrices are independent, so distinct first-order residual directions become orthogonal in normalized inner product as $n \to \infty$, and the first-degree contribution of a single cycle has RMS scale $\sqrt{L}\,c$. Reusing the same stack over N loops adds this contribution coherently, and the assembled linear increment is $Nc\sum_{l=1}^{L} \boldsymbol{W}_0^l \boldsymbol{U}$. Hence

$$\left\| Nc \sum_{l=1}^{L} \boldsymbol{W}_0^l \boldsymbol{U} \right\|_{\mathrm{rms}}^2 \xrightarrow{\mathbb{P}} N^2 L c^2 = a^2 N^{2-2\alpha_N} L^{1-2\alpha_L}.$$

Writing $\beta_{N,L} = N^{1-\alpha_N} L^{1/2-\alpha_L}$, the linear residual contribution has scale $N\sqrt{L}\,c = a\beta_{N,L}$. Thus the initialized residual stream grows with L when $\alpha_L < 1/2$, independently of N. When $\alpha_L = 1/2$ it grows with N if $\alpha_N < 1$, stays of order one if $\alpha_N = 1$, and vanishes toward the input $\boldsymbol{U}$ if $\alpha_N > 1$. When $\alpha_L > 1/2$ it vanishes as $L \to \infty$ for every α_N. Since the limit $L \to \infty$ is taken before $N \to \infty$, Desideratum 1 therefore needs (i) $\alpha_L > \frac{1}{2}$, or (ii) $\alpha_L = \frac{1}{2}$ and $\alpha_N \geq 1$. This keeps the distance between the hidden state and the input bounded at initialization but does not require it to remain nonzero.

One gradient step. Write $\Delta \boldsymbol{x}_0^{r,l}(\zeta) = \boldsymbol{x}_1^{r,l}(\zeta) - \boldsymbol{x}_0^{r,l}(\zeta)$, and define $\Delta \boldsymbol{h}_0^{r,l}$ and Δf_0 analogously, always at the same input. We now work in the initialization-stable regime above.

An SGD step changes each entry of $\boldsymbol{W}^l$ by order $s_{N,L}/n$, where $s_{N,L} = N^{-\gamma_N} L^{-\gamma_L}$. The first block in the first loop receives the raw input $\boldsymbol{U}\zeta$, giving the direct response $\Delta \boldsymbol{h}_0^{1,1}(\zeta) = \Delta \boldsymbol{W}_0^1 \boldsymbol{U}\zeta$. This response has RMS scale $s_{N,L}$. Because we take $L \to \infty$ before $N \to \infty$, boundedness of the first-block preactivation therefore requires (i) $\gamma_L > 0$, or (ii) $\gamma_L = 0$ and $\gamma_N \geq 0$.

Each updated block is then applied N times and multiplied by c at every application. The gradient of every block is aligned with the readout, so these $N \times L$ contributions add coherently in the scalar output rather than as orthogonal hidden-state directions. The aggregate response $|\Delta f_0(\zeta)|$ therefore has scale $NLc\, s_{N,L} = a\rho_{N,L}$, where $\rho_{N,L} = N^{1-\alpha_N-\gamma_N} L^{1-\alpha_L-\gamma_L}$. The first-step output response blows up with L if $1 - \alpha_L - \gamma_L > 0$, vanishes with L if $1 - \alpha_L - \gamma_L < 0$, and, after the L-limit, blows up or vanishes with N according to the sign of $1 - \alpha_N - \gamma_N$. Desideratum 2 asks for a uniformly bounded hidden-state update and an order-one output response, so both exponents must satisfy $\gamma_L = 1 - \alpha_L$ and $\gamma_N = 1 - \alpha_N$, respectively.

Substituting $\gamma_L = 1 - \alpha_L$ and $\gamma_N = 1 - \alpha_N$ back into the local-update bound, and keeping the initialization-stable regime, yields the stable and nontrivial region (the shaded area in Figure 8):

$$\gamma_L = 1 - \alpha_L, \qquad \gamma_N = 1 - \alpha_N, \qquad \begin{cases} \alpha_L = \frac{1}{2}, & \alpha_N \geq 1, \\ \frac{1}{2} < \alpha_L < 1, & \alpha_N \in \mathbb{R}, \\ \alpha_L = 1, & \alpha_N \leq 1. \end{cases} \tag{B3}$$

Desiderata 1 and 2 therefore leave a region rather than a unique point.

To select a point inside this region, we additionally ask for the largest depthwise variation that independent block directions can support, as in Depth-μP (Yang et al., 2024a). This is the layer analogue of the order-one hidden-state change at a fixed loop fraction in Section 3.2.

Desideratum 3 (Maximal feature diversity)**.** At initialization and on a fixed nonzero input, for every fixed $\lambda \in [0, 1)$ and sufficiently small $0 < \varepsilon \leq 1 - \lambda$, the rescaled hidden-state change across an ε-fraction of the first stack should satisfy $N\left\| \boldsymbol{x}_0^{1,\lfloor(\lambda+\varepsilon)L\rfloor} - \boldsymbol{x}_0^{1,\lfloor\lambda L\rfloor} \right\|_{\mathrm{rms}} = \Theta(\sqrt{\varepsilon})$ after the ordered network limits and then as $\varepsilon \downarrow 0$.

Across εL independent blocks the hidden-state increment has RMS scale $c\sqrt{\varepsilon L}$. Multiplying by the outer factor N in Desideratum 3 therefore gives $Nc\sqrt{\varepsilon L} = a\beta_{N,L}\sqrt{\varepsilon}$. The desideratum asks for this quantity to be $\Theta(\sqrt{\varepsilon})$ as $\varepsilon \downarrow 0$, hence $\beta_{N,L} = \Theta(1)$. In the ordered limits, this forces $\alpha_L = 1/2$ and then $\alpha_N = 1$. The stable-and-nontrivial relations then give $\gamma_L = 1/2$ and $\gamma_N = 0$, and select $(\alpha_N, \alpha_L; \gamma_N, \gamma_L) = \left(1, \frac{1}{2}; 0, \frac{1}{2}\right)$, which we call Loop-$\mu$P. The formal classification and the feature properties in both the depth and loop directions are stated in Appendix B.3.

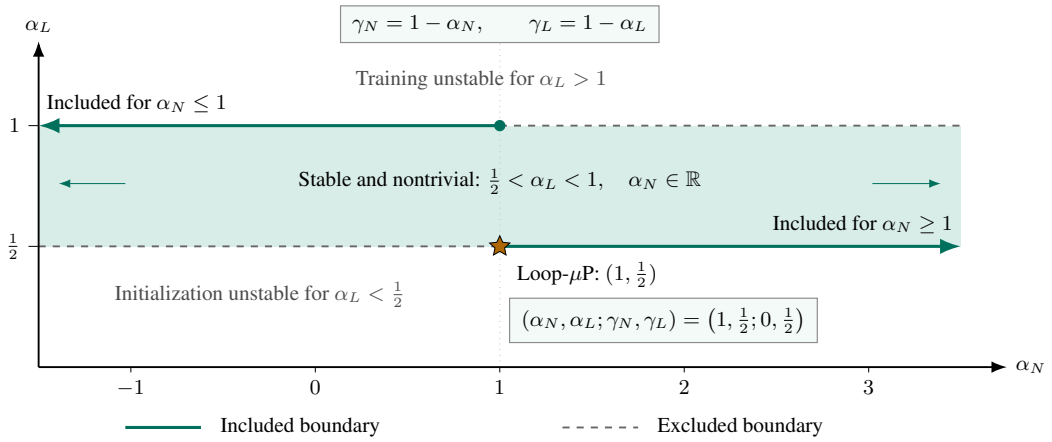


Figure 8: Loop-depth classification on the two-dimensional slice $\gamma_N = 1 - \alpha_N$, $\gamma_L = 1 - \alpha_L$, from Claims 4–6. The shaded strip shows the interior of the stable and nontrivial region; solid boundaries are included and dashed boundaries are excluded. The star marks Loop-μP, distinguished by its nondegenerate initialization loop trajectory and attained rescaled depth-diversity exponent $\kappa_L = 1/2$.

B.3 Classification of loop-depth parameterizations

We now parameterize the loop-depth model by the four exponents $(\alpha_N, \alpha_L; \gamma_N, \gamma_L)$. Because the limit $L \to \infty$ is taken before $N \to \infty$, a power $N^{u_N} L^{u_L}$ is bounded in the ordered limit exactly when $u_L < 0$, or when $u_L = 0$ and $u_N \le 0$. We denote this lexicographic condition by $(u_L, u_N) \preceq_{\text{lex}} (0, 0)$, comparing the depth component first and using the loop component only to break ties. We use $\prec_{\text{lex}}$ for the strict order and $\succeq_{\text{lex}}, \succ_{\text{lex}}$ for the corresponding reverse orders. We use the input batches, training routines, and remaining conventions from Appendix A.1. We first characterize initialization stability, training stability, and nontriviality.

Claim 4. A parameterization is stable at initialization if and only if $(1/2 - \alpha_L,\ 1 - \alpha_N) \preceq_{\text{lex}} (0, 0)$.

Equivalently, $\alpha_L > 1/2$ is sufficient for every α_N. On the boundary $\alpha_L = 1/2$, the remaining condition is $\alpha_N \ge 1$, matching the single-matrix residual-branch result in Claim 1.

Claim 5. Consider a parameterization that is stable at initialization. We characterize training stability and nontriviality separately:

(i) It is stable during training if and only if both

$$(-\gamma_L, -\gamma_N) \preceq_{\text{lex}} (0, 0),$$
$$(1 - \alpha_L - \gamma_L,\ 1 - \alpha_N - \gamma_N) \preceq_{\text{lex}} (0, 0).$$

(ii) It is nontrivial if and only if $(1 - \alpha_L - \gamma_L,\ 1 - \alpha_N - \gamma_N) \succeq_{\text{lex}} (0, 0)$.

In particular, when $(1 - \alpha_L - \gamma_L,\ 1 - \alpha_N - \gamma_N) \prec_{\text{lex}} (0, 0)$, no training routine changes the output in the limit.

The first lexicographic condition controls the local-update scale $s_{N,L}$, while the second controls the aggregate-response scale $\rho_{N,L}$. Combining the two claims gives the formal stable-and-nontrivial region: $\gamma_L = 1 - \alpha_L$ and $\gamma_N = 1 - \alpha_N$, with (α_L, α_N) lying in

$$\left(\frac{1}{2}, 1\right) \preceq_{\text{lex}} (\alpha_L, \alpha_N) \preceq_{\text{lex}} (1, 1).$$

The resulting region contains infinitely many points, as shown in Figure 8. We therefore need an additional criterion that distinguishes feature variation along the loop and depth directions.

Definition 6 measures variation along a single loop coordinate. Here the repeated stack has both loop and depth coordinates, and a within-cycle depth increment carries the loop factor $1/N$. We therefore introduce an unscaled loop increment and an N-rescaled depth increment, and apply the same exponent convention to both.

Definition 7 (Loop-depth feature increments)**.** We write $\boldsymbol{y}_t^0 = \boldsymbol{U}\zeta$ and $\boldsymbol{y}_t^r = \boldsymbol{x}_t^{r,L}$ for the hidden state after r complete cycles. We also write $r_N(\rho) = 1 + \lfloor \rho(N-1) \rfloor$ for the loop at fraction ρ. For $\rho, \lambda \in [0, 1)$ and sufficiently small $\varepsilon > 0$, we define the loop increment and the rescaled depth increment by

$$D_{N,t}(\rho, \varepsilon) = \left\| \boldsymbol{y}_t^{\lfloor (\rho+\varepsilon)N \rfloor} - \boldsymbol{y}_t^{\lfloor \rho N \rfloor} \right\|_{\mathrm{rms}},$$
$$D_{L,t}(\rho, \lambda, \varepsilon) = N \left\| \boldsymbol{x}_t^{r_N(\rho), \lfloor (\lambda+\varepsilon)L \rfloor} - \boldsymbol{x}_t^{r_N(\rho), \lfloor \lambda L \rfloor} \right\|_{\mathrm{rms}}.$$

On a nondegenerate evaluation batch, we define κ_N and κ_L by applying the exponent convention in Definition 6 to $D_{N,t}$ and $D_{L,t}$, respectively, taking the network limits before $\varepsilon \downarrow 0$. The factor N in $D_{L,t}$ compensates for the loop factor $1/N$, without which a change within one cycle would vanish in the loop limit.

Claim 6 (Loop-μP)**.** Within the stable and nontrivial region, the only parameterization for which the initialization loop increment is nondegenerate and the rescaled depth-diversity exponent attains $\kappa_L = 1/2$ is

$$(\alpha_N, \alpha_L; \gamma_N, \gamma_L) = \left(1, \frac{1}{2}; 0, \frac{1}{2}\right).$$

We call this parameterization **Loop-μP** (marked in Figure 8). For every SGD routine, the parameterization is faithful. Moreover, for a sufficiently small $\eta > 0$, a first step on a nonzero training input with a nonzero loss derivative changes both the hidden states and the preactivations by order one at every positive loop fraction. The gradient of each hidden matrix is uniformly $\mathcal{O}(n^{-1}L^{-1/2})$. At initialization, the coordinates of the first-step gradients have the sharper scale $\Theta(n^{-1}L^{-1/2})$. For every SGD routine, fixed training step, and nondegenerate evaluation batch, and for fixed $\rho, \lambda \in [0, 1)$,

$$D_{N,t}(\rho, \varepsilon) = \Theta(\varepsilon), \qquad \kappa_N = 0,$$
$$D_{L,t}(\rho, \lambda, \varepsilon) = \Theta(\sqrt{\varepsilon}), \quad \kappa_L = \frac{1}{2}.$$

$\varepsilon > 0$ keeps the positions within range, and the width, depth, and loop limits are taken before $\varepsilon \downarrow 0$. Without the factor N, the depth increment vanishes in the loop limit.

The two directions therefore have different path regularity. The loop trajectory has smooth $\Theta(\varepsilon)$ increments, whereas the N-rescaled within-cycle depth profile has Brownian-scale $\Theta(\sqrt{\varepsilon})$ increments and is rough in this increment-scaling sense. The corresponding unrescaled depth increment still vanishes in the loop limit.

B.4 From Loop-μP to the Transformer recipe

We now translate the distinguished point into the branch multiplier and optimizer update rule. At $(\alpha_N, \alpha_L; \gamma_N, \gamma_L) = \left(1, \frac{1}{2}; 0, \frac{1}{2}\right)$, the branch multiplier is

$$c = aN^{-\alpha_N} L^{-\alpha_L} = \frac{a}{N\sqrt{L}}.$$

For SGD, combining the gradient rescaling with the update coefficient gives the effective coefficient of the raw gradient:

$$\eta N^{\alpha_N - 1 - \gamma_N} L^{\alpha_L - \gamma_L} = \eta.$$

For Adam and other entrywise update maps, the rescaled gradients remain as inputs to Q_t, while the coefficient outside Q_t is

$$\eta n^{-1} N^{-\gamma_N} L^{-\gamma_L} = \frac{\eta}{n\sqrt{L}}.$$

At $\alpha_L = 1/2$, expressing width and depth relative to the base model gives the factors $N^{-1}m_L^{-1/2}$ and $m_n^{-1}m_L^{-1/2}$ in the residual and hidden learning-rate rows of Table 1, respectively. The loop-independent learning-rate coefficient is consistent with the shared gradient accumulating N contributions, each carrying the loop factor $1/N$.

C Additional Experimental Results

This appendix collects the results that support the experiments of Section 4, in the order in which the main text uses them.

C.1 One-step response over the first ten steps

Section 4.1 measures the change of the residual stream after the first optimizer step. Figure 9 extends this measurement to the first ten optimizer steps, with one panel per exponent α_N and one curve per step. The pattern of the first step holds at every step. Under Loop-μP the update stays within a factor of about four over $N = 1$–64 and levels off near $N = 32$, while the unscaled and square-root updates keep growing with N.

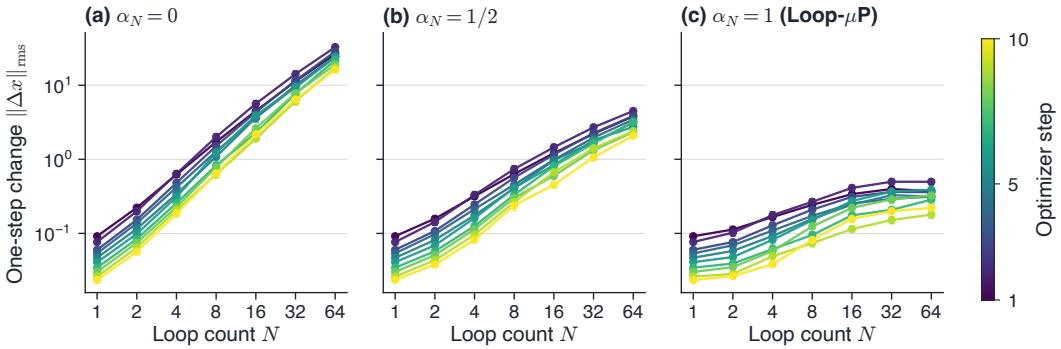


Figure 9: **The one-step pattern of Figure 3 (right) holds over the first 10 optimizer steps.** Each panel fixes one exponent α_N and plots the RMS change of the final residual stream against the loop count N, with one curve per optimizer step t=1, . . . , 10 (viridis ramp, mean ± standard error over 10 seeds). All settings share the base learning rate 10^{-4}. The unscaled and square-root updates grow with N at every step (panels (a) and (b)), while the Loop-μP update stays within a factor of about four over N=1–64 and levels off near N=32 (panel (c)).

C.2 Post-training cosine similarity at a smaller loop count

Section 4.3 shows that the increments of different passes stay aligned after training at $N = 8$. Figure 10 repeats the measurement for the $L \in \{12, 24, 48\}$ models trained at $N = 4$. Adjacent increments in the first three passes have cosine similarities between 0.60 and 0.70 at all three depths. The increment of the final pass is again weakly or negatively correlated with the others, most strongly at $L = 12$.

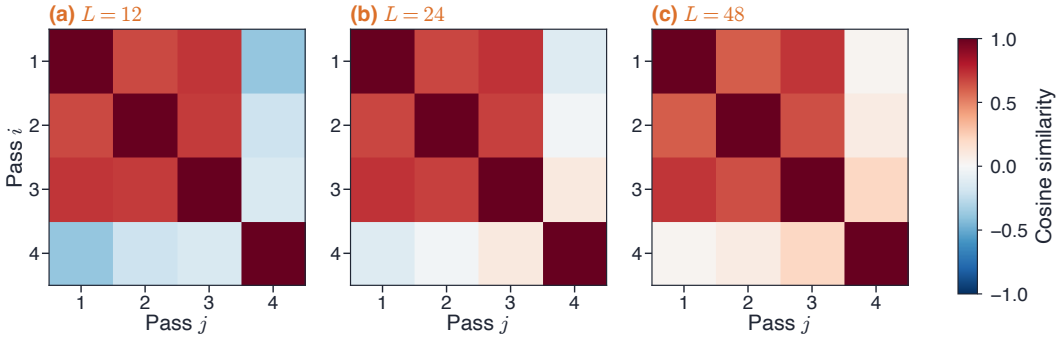


Figure 10: **Cosine similarity between pass increments after full training** (N=4). Companion to Figure 5 (N=8). Early and middle passes remain positively correlated across all three depths. Pairs with the final pass are weaker and can be near zero or negative, especially at L=12, matching the pattern in Figure 5.

C.3 Post-training layer increments of shallower untied models

Section 4.3 uses the per-layer increments of a trained untied $L = 48$ model as a control for the pass increments of looped models. Figure 11 repeats this measurement for the trained untied $L = 12$ and $L = 24$ models, plotting the cosine similarity between every pair of per-layer increments, averaged over many sequences. The pattern matches $L = 48$. Early layers are positively correlated with one another, but the similarity fades with layer distance and turns negative for the last layers.

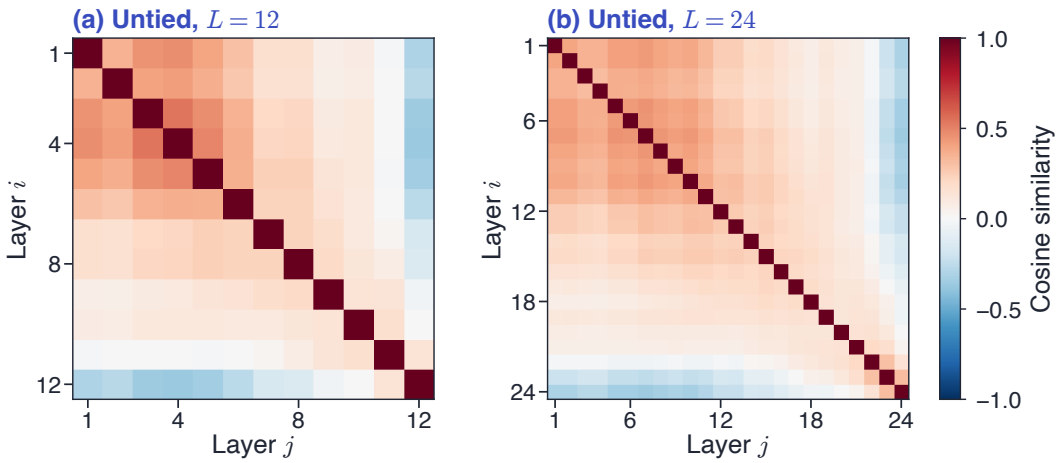


Figure 11: **Layer increments of trained untied models align mainly with nearby layers.** Cosine similarity between the per-layer residual-stream increments of trained untied models with L=12 (a) and L=24 (b), after full training. Companion to the untied L=48 model in panel (a) of Figure 5.

D LIMITATIONS

Following Yang et al. (2024a), our theory analyzes a mean-centered MLP residual block and proves its results for linear networks, and its extension to nonlinear activations and Adam relies on the assumptions of Appendix I. Future work could remove these assumptions and extend the analysis to attention and normalization layers, which would bring the theory closer to the Transformers used in practice. Limited by our compute budget, our largest model has 438M parameters and is trained on 10B tokens, and we leave the study of Loop-μP at larger scales to future work.

E EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

This appendix lists the configurations behind the experiments of Section 4 and the instance of Table 1 that they use.

E.1 LANGUAGE MODELING EXPERIMENT CONFIGURATION

Table 2 lists the model configurations and training hyperparameters for the language-modeling experiments of Section 4.2.

Table 2: **Model configurations for the language-modeling experiments.** "KV" denotes key/value attention heads (each row uses 12 attention heads and 12 key/value heads, i.e. no GQA grouping). All models are decoder-only Llama-style pre-norm Transformers with RMSNorm and the Llama 3 tokenizer vocabulary (Vaswani et al., 2017; Touvron et al., 2023; Llama Team, 2024; Xiong et al., 2020; Zhang & Sennrich, 2019). The token embedding and tied LM head are outside the loop (Press & Wolf, 2017), while the whole Transformer stack is repeated N times. Training uses FineWeb-Edu for 20,000 optimizer steps (10B tokens, 0.5M-token global batch), AdamW (Loshchilov & Hutter, 2019) with $(\beta_1, \beta_2) = (0.9, 0.95)$ and weight decay ω_0 (see Appendix E.3), and a warmup-stable-linear-decay schedule (Wen et al., 2024) with 500 warmup steps and a final 1,000-step decay.

Layers L	Loop counts N	d_{model}	Attn/KV heads	Head dim	MLP dim	Vocab	Params (tied)
12	$\{1, 2, 4, 8\}$	768	12/12	64	2048	128256	183.5M
24	$\{1, 2, 4, 8\}$	768	12/12	64	2048	128256	268.4M
48	$\{1, 2, 4, 8\}$	768	12/12	64	2048	128256	438.3M

E.2 LOOP-μP PARAMETERIZATION USED IN THE EXPERIMENTS

Table 3 instantiates Table 1 for our experiments. We keep the width fixed ($m_n = 1$) and use $L_{\text{base}} = 12$, so the depth multiplier is $m_L = L/12$. We write η_0 for the base learning rate at $L = 12$,

σ_0^2 for the base initialization variance, ω_0 for the base AdamW weight decay, and ϵ_{opt} for the base AdamW numerical epsilon.

Table 3: **Loop-μP as used in our experiments.** The loop factor N^{-1} appears only in the attention and MLP residual branches, while the factor $m_L^{-1/2}$ with $m_L = L/12$ scales both these branches and the learning rates of the Transformer layers. Learning rates depend on L but not on N.

Quantity	Scaling rule
Token embedding init. variance	σ_0^2
Token embedding LR	η_0
Pre-LN init. variance	σ_0^2
Pre-LN LR	$\eta_0\, m_L^{-1/2}$
Transformer hidden init. variance	σ_0^2
Transformer hidden LR	$\eta_0\, m_L^{-1/2}$
Transformer hidden bias LR	$\eta_0\, m_L^{-1/2}$
Transformer hidden AdamW weight decay	ω_0
Attention residual branch	$\boldsymbol{X} + a\, N^{-1} m_L^{-1/2} \cdot \mathrm{MHA}(\mathrm{LN}(\boldsymbol{X}))$
MLP residual branch	$\boldsymbol{Z} + a\, N^{-1} m_L^{-1/2} \cdot \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Z}))$
Final-LN init. variance	σ_0^2
Final-LN LR	η_0
Unembedding init. variance	σ_0^2
Unembedding LR	η_0
Unembedding forward map	$\boldsymbol{X}\boldsymbol{W}_{\text{unemb}}^{\mathsf{T}}$
AdamW ϵ_{opt} for residual blocks	$\epsilon_{\text{opt}}\, m_L^{-1/2}$
AdamW ϵ_{opt} for embedding and unembedding	ϵ_{opt}

E.3 Hyperparameter values and sweep configuration

Table 4 lists the numerical values used to instantiate the parameterization of Appendix E.2 for the language-modeling experiments (Section 4.2) and the learning-rate sweep behind Figure 4.

Table 4: Numerical values used in the language-modeling experiments. Symbols refer to the parameterization in Table 3. Initialization uses a truncated normal with standard deviation σ_0.

Quantity	Symbol	Value
Residual branch constant	a	1
Base learning rate at L=12	η_0	1.25×10^{-3}
Initialization standard deviation	σ_0	0.02
AdamW weight decay	ω_0	0.1
AdamW numerical epsilon	ϵ_{opt}	10^{-8}
LR sweep grid (base η_0)	—	$\{5,\ 7.5,\ 10,\ 12.5,\ 15,\ 20,\ 30,\ 40\} \times 10^{-4}$
Divergence criterion	—	final validation loss > 4
Random seed	—	0 (single run)
Validation split	—	FineWeb-Edu held-out

F Notation and Mathematical Tools

This appendix fixes the notation and collects the probabilistic tools used in the proofs of Appendices G and H. We write $\mathbb{R}$ for the real numbers and $\mathbb{R}^d$ for d-dimensional Euclidean space, let $\mathbf{1}_d \in \mathbb{R}^d$ denote the all-ones vector, and let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix. The notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian law with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathbb{1}_{\{\cdot\}}$ denotes an indicator function. We write $\mathbb{E}$ and $\mathbb{P}$ for expectation and probability, and Tr for the trace. We write $\|\cdot\|$ for the Euclidean or Hilbert norm and use $\|\boldsymbol{z}\|_{\text{rms}} = n^{-1/2}\|\boldsymbol{z}\|$ and $\|\boldsymbol{X}\|_{\text{rms}} = n^{-1/2}\|\boldsymbol{X}\|_{\text{F}}$ for normalized finite-width vector

and batch norms. The macros $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_{\mathrm{F}}$, and $\|\cdot\|_{\mathrm{HS}}$ denote the induced operator, Frobenius, and Hilbert–Schmidt norms, respectively. We use $\mathcal{O}(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$, and $\Theta(\cdot)$ for standard asymptotic notation. We write $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\mathbb{P}}$, and $\xrightarrow{\mathcal{D}}$ for almost sure convergence, convergence in probability, and convergence in distribution, respectively.

Lemma 1 (Isserlis's Theorem (Isserlis, 1918)). *Let $\boldsymbol{x} = (x_1, \dots, x_m)$ be a centered Gaussian random vector. If m is odd then $\mathbb{E}[x_1 \cdots x_m] = 0$. If $m = 2\ell$ then*

$$\mathbb{E}[x_1 \cdots x_{2\ell}] = \sum_{\pi \in \mathcal{P}_2(2\ell)} \prod_{\{a,b\} \in \pi} \mathbb{E}[x_a x_b],$$

where $\mathcal{P}_2(2\ell)$ is the set of pairings of $\{1, \dots, 2\ell\}$. The same formula follows by iterating $\mathbb{E}[gF(g)] = \mathbb{E}[F'(g)]$ for a standard normal g and a polynomial F.

Proof. Induct on m. For a jointly centered Gaussian vector, Gaussian integration by parts gives $\mathbb{E}[x_1 F(\boldsymbol{x})] = \sum_a \mathbb{E}[x_1 x_a]\, \mathbb{E}[\partial_a F(\boldsymbol{x})]$: write $\boldsymbol{x}$ as a deterministic linear image of independent standard Gaussians and apply the one-dimensional identity to each independent coordinate. Applied with $F = x_2 \cdots x_m$, each derivative selects one factor x_a, giving $\mathbb{E}[x_1 \cdots x_m] = \sum_{a=2}^{m} \mathbb{E}[x_1 x_a]\, \mathbb{E}\left[\prod_{b \neq 1,a} x_b\right]$; iterating expands the expectation over all pairings. For odd m one unpaired centered factor remains, giving zero. □

For completeness, if $\boldsymbol{u}, \boldsymbol{v}$ are independent standard Gaussian vectors and $\boldsymbol{M}$ is deterministic, Isserlis's formula also gives

$$\begin{aligned} \mathbb{E}[\boldsymbol{u}^{\mathsf{T}} \boldsymbol{M} \boldsymbol{u}] &= \mathrm{Tr}(\boldsymbol{M}), \\ \mathrm{Var}(\boldsymbol{u}^{\mathsf{T}} \boldsymbol{M} \boldsymbol{u}) &= 2 \left\| \frac{\boldsymbol{M} + \boldsymbol{M}^{\mathsf{T}}}{2} \right\|_{\mathrm{F}}^2, \\ \mathbb{E}[(\boldsymbol{u}^{\mathsf{T}} \boldsymbol{M} \boldsymbol{v})^2] &= \|\boldsymbol{M}\|_{\mathrm{F}}^2. \end{aligned}$$

Write $\mathrm{O}(n) = \left\{\boldsymbol{Q} \in \mathbb{R}^{n \times n} : \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{Q} = \boldsymbol{I}_n\right\}$ for the standard orthogonal group. We write $\boldsymbol{Q} \sim \mathrm{Haar}(\mathrm{O}(n))$ for a uniformly random orthogonal matrix. A Haar rotation is left multiplication by an independent matrix with this law. An orthonormal k-frame is a matrix $\boldsymbol{E} \in \mathbb{R}^{n \times k}$ satisfying $\boldsymbol{E}^{\mathsf{T}} \boldsymbol{E} = \boldsymbol{I}_k$; a Haar k-frame is the first k columns of a Haar orthogonal matrix, equivalently a uniformly distributed frame invariant under left multiplication by every fixed element of $\mathrm{O}(n)$. The lemma below uses only this invariance, the Gaussian frame construction, the rotated-row Gaussian limit, and the coordinate moment identities.

Lemma 2 (Diaconis & Freedman (1987)). *Let $k, r \geq 1$ be fixed and let $\boldsymbol{Q} \sim \mathrm{Haar}(\mathrm{O}(n))$, $n \geq \max(k, r)$.*

(i) *A Haar k-frame in $\mathbb{R}^n$ has the law $\boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1/2}$, where $\boldsymbol{Y}$ is an $n \times k$ matrix of i.i.d. standard normals. The $r \times k$ matrix formed by the first r rows of $\sqrt{n}\,\boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1/2}$ converges in distribution to a matrix of i.i.d. standard normals.*

(ii) *Let $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$ be deterministic and $\boldsymbol{\Sigma} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}/n$. The first r rows of $\boldsymbol{Q}\boldsymbol{Z}$ converge in distribution to r independent $\mathcal{N}(0, \boldsymbol{\Sigma})$ rows in $\mathbb{R}^k$. The convergence is uniform for $\boldsymbol{\Sigma}$ in any bounded set of PSD $k \times k$ matrices, and includes singular $\boldsymbol{\Sigma}$.*

(iii) *For every $\boldsymbol{z} \in \mathbb{R}^n$ and every coordinate i,*

$$\mathbb{E}_{\boldsymbol{Q}}\left[|(\boldsymbol{Q}\boldsymbol{z})_i|^2\right] = \|\boldsymbol{z}\|_{\mathrm{rms}}^2, \qquad \mathbb{E}_{\boldsymbol{Q}}\left[|(\boldsymbol{Q}\boldsymbol{z})_i|^4\right] = \frac{3n}{n+2}\,\|\boldsymbol{z}\|_{\mathrm{rms}}^4.$$

Proof. For (i), left multiplication of $\boldsymbol{Y}$ by a fixed $\boldsymbol{R} \in \mathrm{O}(n)$ preserves its law, so the law of $\boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1/2}$ is invariant under left multiplication; the unique such uniform law on the set of orthonormal k-frames is the law of a Haar k-frame. Its first r rows multiplied by $\sqrt{n}$ equal $\boldsymbol{Y}_r(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}/n)^{-1/2}$, where $\boldsymbol{Y}_r$ is the top $r \times k$ block of $\boldsymbol{Y}$. The law of large numbers gives $\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}/n \xrightarrow{\mathbb{P}} \boldsymbol{I}_k$, and the continuous-mapping theorem yields the asserted convergence. For (ii), extend the column space of $\boldsymbol{Z}$ to that of an orthonormal k-frame $\boldsymbol{E}$ and write $\boldsymbol{Z} = \sqrt{n}\,\boldsymbol{E}\boldsymbol{A}$ with $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} = \boldsymbol{\Sigma}$; the polar factorization gives $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{\Sigma}^{1/2}$ for an orthogonal $\boldsymbol{R}$ (zero columns

allowed). Then $\boldsymbol{QZ} = \sqrt{n}\,(\boldsymbol{QER})\,\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{QER}$ is again a Haar k-frame, so (i) applies. The error is uniform for $\boldsymbol{\Sigma}$ in a bounded set since $\boldsymbol{\Sigma}$ enters only through the bounded matrix $\boldsymbol{\Sigma}^{1/2}$, and no invertibility of $\boldsymbol{\Sigma}$ is used. For (iii), $\boldsymbol{Qz}/\|\boldsymbol{z}\|$ is uniform on the unit sphere, hence equal in law to $\boldsymbol{y}/\|\boldsymbol{y}\|$ for standard Gaussian $\boldsymbol{y}$ in $\mathbb{R}^n$. The direction $\boldsymbol{y}/\|\boldsymbol{y}\|$ and the radius $\|\boldsymbol{y}\|$ are independent, so $\mathbb{E}\left[|(\boldsymbol{Qz})_i|^2\right] = \|\boldsymbol{z}\|^2\,\mathbb{E}\left[y_i^2\right]/\,\mathbb{E}\left[\|\boldsymbol{y}\|^2\right] = \|\boldsymbol{z}\|^2/n = \|\boldsymbol{z}\|_{\mathrm{rms}}^2$ and $\mathbb{E}\left[|(\boldsymbol{Qz})_i|^4\right] = \|\boldsymbol{z}\|^4\,\mathbb{E}\left[y_i^4\right]/\,\mathbb{E}\left[\|\boldsymbol{y}\|^4\right] = 3\|\boldsymbol{z}\|^4/(n(n+2)) = \frac{3n}{n+2}\|\boldsymbol{z}\|_{\mathrm{rms}}^4$. □

For a family $\{Z_\theta : \theta \in \mathcal{I}\}$, uniform anti-concentration at zero means

$$\lim_{\varepsilon \downarrow 0} \sup_{\theta \in \mathcal{I}} \mathbb{P}\left(|Z_\theta| \le \varepsilon\right) = 0. \tag{F1}$$

Together with upper-tail tightness, this is the lower-scale condition in the weak Θ statements below. The random-rotation moment identities transfer normalized Gram bounds to this condition. More precisely, let $\mathcal{G}$ be a sigma-algebra, let σ^2 be $\mathcal{G}$-measurable, and suppose $Z \mid \mathcal{G} \sim \mathcal{N}(0, \sigma^2)$. If $\sigma^2 \ge m > 0$ almost surely, then, for every $\varepsilon > 0$,

$$\mathbb{P}\left(|Z| \le \varepsilon \mid \mathcal{G}\right) \le \sqrt{\frac{2}{\pi}}\,\frac{\varepsilon}{\sqrt{m}} \quad \text{almost surely.} \tag{F2}$$

If instead $\sigma^2 > 0$ almost surely without a deterministic lower bound, then for every $m > 0$,

$$\mathbb{P}\left(|Z| \le \varepsilon\right) \le \mathbb{P}\left(\sigma^2 < m\right) + \sqrt{\frac{2}{\pi}}\,\frac{\varepsilon}{\sqrt{m}};$$

letting first $\varepsilon \downarrow 0$ and then $m \downarrow 0$ proves Eq. (F1).

G Proofs for the Infinite-Loop Limit in the Linear Case

This appendix proves Claims 1–3 of Appendix A.2 and Claims 2′ and 3′ of Appendix A.3 for linear networks trained with SGD; the calculation in Appendix G.4 is diagnostic rather than part of the strict classification. The main results are the classification in Appendix G.2 and the exact first-step response in Proposition 8, valid for every fixed finite $k \ge 1$. They establish stability, nontriviality within the stable region, and learning of the hidden state and every internal feature. Lemma 7 supplies the all-routine lower bounds for faithfulness and loop diversity. We first prove these results in Appendices G.1 and G.2, then strengthen the nontriviality classification for $k = 1$ in Appendix G.3. The $k = 2$ calculation in Appendix G.4 shows why that strengthening cannot hold for all fixed block depths.

Model. The frozen input/readout matrices and hidden matrices are mutually independent, with $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{V}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-2})$, and $(\boldsymbol{W}_0^j)_{ii'} \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-1})$ for $1 \le j \le k$. For a scalar input ζ, the identity-activation network is

$$\begin{aligned}
\boldsymbol{x}_t^0(\zeta) &= \boldsymbol{U}\zeta, \\
\boldsymbol{h}_t^{r,j}(\zeta) &= \boldsymbol{W}_t^j \cdots \boldsymbol{W}_t^1 \boldsymbol{x}_t^{r-1}(\zeta), \qquad 1 \le j \le k, \\
\boldsymbol{x}_t^r(\zeta) &= \boldsymbol{x}_t^{r-1}(\zeta) + c_N \boldsymbol{h}_t^{r,k}(\zeta), \qquad 1 \le r \le N, \\
f_t(\zeta) &= \boldsymbol{V}^{\mathsf{T}} \boldsymbol{x}_t^N(\zeta).
\end{aligned}$$

The input and readout matrices are frozen, while all hidden matrices are trainable and shared across the N loops. Each matrix uses its complete gradient over all loops. With

$$\boldsymbol{G}_{j,t} = \nabla_{\boldsymbol{W}^j} \ell_t(f_t(\xi_t)), \qquad \widetilde{\boldsymbol{G}}_{j,t} = nN^{\alpha_N - 1}\boldsymbol{G}_{j,t},$$

the simultaneous SGD update is

$$\boldsymbol{W}_{t+1}^j = \boldsymbol{W}_t^j - \eta n^{-1} N^{-\gamma_N} \widetilde{\boldsymbol{G}}_{j,t} = \boldsymbol{W}_t^j - \eta N^{\alpha_N - 1 - \gamma_N} \boldsymbol{G}_{j,t}.$$

Scales and limits. The constants $k, a, \eta > 0$ and the finite training horizon T are fixed. Width $n \to \infty$ precedes the loop limit $N \to \infty$. Put

$$c_N = aN^{-\alpha_N}, \qquad \beta_N = N^{1-\alpha_N}, \qquad s_N = N^{-\gamma_N}, \qquad \rho_N = \beta_N s_N.$$

Thus $Nc_N = a\beta_N$. Here c_N is the branch multiplier, β_N measures the change from the input at initialization, s_N the direct effect of an update, and ρ_N its effect on the output.

Training routine. The linear routines below use a fixed horizon, bounded inputs and loss derivatives, and simultaneous SGD updates. In singleton notation, $\Xi_t = (\xi_t)$, $\chi_t = \ell_t'(f_t(\xi_t))$, and $d_t = \eta\chi_t\xi_t$. Estimates are first uniform on compact control sets, then transferred to tight realized controls by truncation and subsequences. First-step lower bounds use $\xi_0 \neq 0$, $\chi_\star \neq 0$, and $d_\star = \eta\chi_\star\xi_0$. A loss derivative continuous and nonzero at zero supplies this signal because the initialization output has width limit zero.

Differences compare the same input. Upper bounds use fixed bounded input batches; lower bounds use $0 < c_\zeta \leq |\zeta_j| \leq C_\zeta < \infty$. Vector and column-stacked batch magnitudes use $\|\cdot\|_{\mathrm{rms}}$. Coordinatewise upper bounds mean tightness along all allowed position sequences; lower bounds additionally require the anti-concentration condition in Eq. (F1), obtained from the random-rotation lemma in Appendix F.

For the linear model, a finite training batch reduces exactly to the scalar controls used below. If $\Xi_t = (\xi_{t,b})_{b=1}^B$, with B fixed independently of network size, and ℓ_t is a differentiable loss of the batch outputs, then $f_t(\xi) = \xi f_t(1)$ gives

$$\nabla_\theta \ell_t = \left(\sum_{b=1}^B \xi_{t,b}\, \partial_b \ell_t\right) \nabla_\theta f_t(1), \qquad d_t = \eta \sum_{b=1}^B \xi_{t,b}\, \partial_b \ell_t.$$

Here θ denotes any trainable matrix or the full parameter list, and each $\partial_b \ell_t$ is evaluated at the current batch outputs. In the singleton formulas below, replace $\chi_t\xi_t$ by d_t/η for a training batch. This gives the same controlled recursions and a tight d_t under the bounded-input and bounded-loss-derivative assumptions. Existential first-step lower bounds may still use the stated singleton signal.

Colored representation. Put $\boldsymbol{u} = \boldsymbol{U}$ and $\boldsymbol{v} = n\boldsymbol{V}$. We record the input and readout in two source copies of a Fock space: a Hilbert space of words, in which each letter marks one matrix application together with its transpose sign. Written in this word basis, expansions of matrix products become word combinatorics, and the only operators we need are *creation*, which prepends a letter, and its adjoint *annihilation*, which contracts the first letter.

Definition 8. For a real Hilbert space H, the full Fock space on signed copies of H is $\mathcal{F}(H \oplus H) = \bigoplus_{m\geq 0}(H \oplus H)^{\otimes m}$: the degree-$m$ summand is the space of length-m words in the letters $(f,+), (f,-)$ for $f \in H$, and the degree-zero summand is the empty word. Left creation $\mathsf{L}(f,\pm)$ prepends $(f,\pm)$ to a word; its adjoint $\mathsf{L}(f,\pm)^*$ contracts the first letter and kills the empty word. On two copies $\mathcal{H}_u \oplus \mathcal{H}_v$, whose empty-word unit vectors $u_\star, v_\star$ are called vacua, define

$$C(f) = \mathsf{L}(f,+) + \mathsf{L}(f,-)^*, \qquad f \in H. \tag{G1}$$

Here $H = \mathbb{R}^k$ with orthonormal basis $e_1, \ldots, e_k$, and

$$C_j = C(e_j), \qquad \|C_j\|_{\mathrm{op}} \leq 2. \tag{G2}$$

Both source copies have orthonormal word bases $e_{w,u}, e_{w,v}$, with letters $(j,+), (j,-)$. Each C_j preserves the source. For $k = 1$, write $C = C_1$; then $C^m u_\star = e_{(1,+)^m,u}$ and $(C^*)^m v_\star = e_{(1,-)^m,v}$.

G.1 SGD TRAINING DYNAMICS

We proceed from finite identities to the colored width limit, then to the coupled-matrix loop limit, and finally to prefix noncancellation and the exact first-step response. At finite width, write $\boldsymbol{R}_{0,t} = \boldsymbol{I}$, $\boldsymbol{R}_{j,t} = \boldsymbol{W}_t^j \cdots \boldsymbol{W}_t^1$, $\boldsymbol{S}_{j,t} = \boldsymbol{R}_{j-1,t}$, $\boldsymbol{T}_{j,t} = \boldsymbol{W}_t^k \cdots \boldsymbol{W}_t^{j+1}$, $\boldsymbol{P}_t = \boldsymbol{R}_{k,t}$, and $\boldsymbol{A}_t = \boldsymbol{I} + c_N \boldsymbol{P}_t$, with empty products interpreted as identities. The Hilbert rank-one map $(z \otimes w)y = z\langle w, y\rangle$ corresponds to $\boldsymbol{z}\boldsymbol{w}^\mathsf{T}/n$ at finite width.

Widthwise limit.

Lemma 3. *For every n, N, t,*

$$\begin{aligned} \boldsymbol{x}_t^r(\zeta) &= \zeta \boldsymbol{A}_t^r \boldsymbol{u}, \\ \boldsymbol{h}_t^{r,j}(\zeta) &= \zeta \boldsymbol{R}_{j,t} \boldsymbol{A}_t^{r-1} \boldsymbol{u}, \\ f_t(\zeta) &= \zeta\, n^{-1} \boldsymbol{v}^\mathsf{T} \boldsymbol{A}_t^N \boldsymbol{u}. \end{aligned} \tag{G3}$$

Let

$$\begin{aligned} \boldsymbol{z}_{j,t,s} &\coloneqq \boldsymbol{S}_{j,t} \boldsymbol{A}_t^s \boldsymbol{u}, \\ \boldsymbol{b}_{j,t,s} &\coloneqq \boldsymbol{T}_{j,t}^\mathsf{T} (\boldsymbol{A}_t^\mathsf{T})^{N-1-s} \boldsymbol{v}. \end{aligned} \tag{G4}$$

Then the exact shared-gradient identities are

$$\begin{aligned} \boldsymbol{G}_{j,t} &= \frac{c_N \chi_t \xi_t}{n} \sum_{s=0}^{N-1} \boldsymbol{b}_{j,t,s} \boldsymbol{z}_{j,t,s}^{\mathsf{T}}, \\ \boldsymbol{W}_{t+1}^{j} - \boldsymbol{W}_{t}^{j} &= -\frac{a d_t N^{-\gamma_N}}{N} \sum_{s=0}^{N-1} \frac{\boldsymbol{b}_{j,t,s} \boldsymbol{z}_{j,t,s}^{\mathsf{T}}}{n}. \end{aligned} \tag{G5}$$

For $\alpha_N \geq 1$, $\gamma_N \geq 0$, each fixed number of training steps, and controls in a compact set, there are bounds independent of n, N on all trained matrices, prefixes, suffixes, residual powers $\boldsymbol{A}_t^r$ for $0 \leq r \leq N$, and the associated normalized vector norms, on the event that the initial operator and vector norms are bounded. In particular,

$$\begin{aligned} \left\| \boldsymbol{W}_{t+1}^{j} - \boldsymbol{W}_{t}^{j} \right\|_{\mathrm{op}} &\leq C N^{-\gamma_N}, \\ \|\boldsymbol{P}_{t+1} - \boldsymbol{P}_t\|_{\mathrm{op}} + \|\boldsymbol{R}_{j,t+1} - \boldsymbol{R}_{j,t}\|_{\mathrm{op}} &\leq C N^{-\gamma_N}, \\ \|\Delta \boldsymbol{x}_t^r(\zeta)\|_{\mathrm{rms}} + |\Delta f_t(\zeta)| &\leq C N^{1-\alpha_N-\gamma_N}, \\ \left\| \Delta \boldsymbol{h}_t^{r,j}(\zeta) \right\|_{\mathrm{rms}} &\leq C N^{-\gamma_N}. \end{aligned} \tag{G6}$$

Constants may depend on the fixed k, a, t, η, the compact control set, and input bounds, but not on n, N, r.

Proof. The recurrence gives Eq. (G3). A perturbation $\delta \boldsymbol{W}^j$ in loop $s+1$ contributes

$$c_N \boldsymbol{T}_{j,t} \, \delta \boldsymbol{W}^j \boldsymbol{S}_{j,t} \boldsymbol{A}_t^s \boldsymbol{u}.$$

Back-propagating the readout through the remaining $N-1-s$ loops and summing over all loops gives the raw gradient in Eq. (G5). Multiplication by the prescribed raw-gradient learning-rate coefficient gives the update formula; no independent-layer replacement is made. For the bounds, suppose every matrix norm at time t is at most $M_t \geq 1$, the initial normalized vector norms are at most D, and $N c_N \leq a$. Then

$$\begin{aligned} \|\boldsymbol{A}_t^r\|_{\mathrm{op}} &\leq e^{a M_t^k}, \\ \|\boldsymbol{z}_{j,t,s}\|_{\mathrm{rms}} &\leq D M_t^{j-1} e^{a M_t^k}, \\ \|\boldsymbol{b}_{j,t,s}\|_{\mathrm{rms}} &\leq D M_t^{k-j} e^{a M_t^k}. \end{aligned}$$

Using $\|\boldsymbol{p}\boldsymbol{q}^{\mathsf{T}}/n\|_{\mathrm{op}} = \|\boldsymbol{p}\|_{\mathrm{rms}} \|\boldsymbol{q}\|_{\mathrm{rms}}$ in Eq. (G5) gives

$$M_{t+1} \leq M_t + a|d_t| D^2 M_t^{k-1} e^{2 a M_t^k}.$$

This finite induction proves the matrix bounds and the first line of Eq. (G6); telescoping the fixed k matrices proves its second line. Finally,

$$\boldsymbol{A}_{t+1}^r - \boldsymbol{A}_t^r = \sum_{s=0}^{r-1} \boldsymbol{A}_{t+1}^{r-1-s} c_N (\boldsymbol{P}_{t+1} - \boldsymbol{P}_t) \boldsymbol{A}_t^s \tag{G7}$$

proves the state bound. Pairing with $\boldsymbol{v}/n$ proves the output bound, while

$$\boldsymbol{R}_{j,t+1} \boldsymbol{A}_{t+1}^{r-1} - \boldsymbol{R}_{j,t} \boldsymbol{A}_t^{r-1} = (\boldsymbol{R}_{j,t+1} - \boldsymbol{R}_{j,t}) \boldsymbol{A}_{t+1}^{r-1} + \boldsymbol{R}_{j,t} (\boldsymbol{A}_{t+1}^{r-1} - \boldsymbol{A}_t^{r-1})$$

proves the internal-state bound. A $1/4$-net of the unit sphere has at most 9^n points, and $\|\boldsymbol{W}_0^j\|_{\mathrm{op}} \leq 2 \max_{\boldsymbol{x},\boldsymbol{y} \text{ in the net}} |\boldsymbol{y}^{\mathsf{T}} \boldsymbol{W}_0^j \boldsymbol{x}|$. Each scalar in this maximum is $\mathcal{N}(0, n^{-1})$, so a union bound gives

$$\mathbb{P}\left(\max_{1 \leq j \leq k} \left\| \boldsymbol{W}_0^j \right\|_{\mathrm{op}} > R \right) \leq 2k \exp(2n \log 9 - nR^2/8). \tag{G8}$$

For fixed k this gives a bounded initialization event with probability tending to one and bounds all fixed moments of the operator norms. Gaussian concentration likewise gives $\|\boldsymbol{u}\|_{\mathrm{rms}}, \|\boldsymbol{v}\|_{\mathrm{rms}} \xrightarrow{\mathbb{P}} 1$ and their fixed-moment bounds. □

Lemma 4. *For fixed mixed words P, Q in the matrices $\boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^k$ and their transposes, write $P(\boldsymbol{W}_0), Q(\boldsymbol{W}_0)$ for their finite-width evaluations and $P(C), Q(C)$ for the corresponding evaluations at $C_1, \dots, C_k$. For $\boldsymbol{e}, \boldsymbol{e}' \in \{\boldsymbol{u}, \boldsymbol{v}\}$,*

$$\frac{1}{n}\langle P(\boldsymbol{W}_0)\boldsymbol{e}, Q(\boldsymbol{W}_0)\boldsymbol{e}'\rangle \xrightarrow{\mathbb{P}} \langle P(C)e_\star, Q(C)e'_\star\rangle. \tag{G9}$$

Any fixed finite computation formed from these vectors by matrix and transpose multiplication, linear combinations, normalized inner products, and normalized rank-one maps has the corresponding Gram and scalar limit, uniformly on compact scalar control sets. For equivariant vector computations and invariant tight controls, every fixed collection of coordinate rows has, along converging control subsequences, a centered Gaussian-mixture limit whose conditional covariance is the limiting Gram matrix; distinct rows are conditionally independent.

Proof. Consider a normalized trace of a mixed word of length $2m$ in the initial matrices and their transposes. Isserlis's formula pairs entries with matching matrix colors and identifies their ordered endpoints. The connected quotient graph has at most m undirected edges and hence at most $m+1$ free vertices, so a pairing contributes n^{V-m-1}. Equality requires a tree with m distinct edges. The trace walk visits every edge twice, once in each direction, so paired entries must have opposite transpose signs and the same matrix color. The walk exhausts each descendant component before returning across its edge, which is exactly noncrossing. Conversely, recursively removing an adjacent pair shows that every noncrossing, opposite-sign, matching-color pairing has $m+1$ free vertices. All other terms lose at least one power of n; odd words have zero expectation. The vacuum expansion of the colored circular operators has the same nested/disjoint pairing rule.

For the variance, expand the product of two normalized traces. Pairings disconnected between the traces give the product of expectations. Every connected pairing with m pairs has at most $m+1$ free vertices, while the two-trace normalization is n^{-m-2}, so its contribution is $\mathcal{O}(n^{-1})$. This proves joint L^2 convergence of the normalized traces.

Conditional on the initial matrices, Gaussian quadratic-form identities give, for any real matrix $\boldsymbol{M}$,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{u}/n \mid \boldsymbol{W}_0] &= \mathrm{Tr}(\boldsymbol{M})/n,\\
\mathrm{Var}(\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{u}/n \mid \boldsymbol{W}_0) &\le 2\|\boldsymbol{M}\|_{\mathrm{op}}^2/n,\\
\mathbb{E}[\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{v}/n \mid \boldsymbol{W}_0] &= 0,\\
\mathbb{E}[|\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{v}/n|^2 \mid \boldsymbol{W}_0] &= \|\boldsymbol{M}\|_{\mathrm{F}}^2/n^2 \le \|\boldsymbol{M}\|_{\mathrm{op}}^2/n.
\end{aligned}$$

Apply these identities to $\boldsymbol{M} = P(\boldsymbol{W}_0)^\mathsf{T}Q(\boldsymbol{W}_0)$. At fixed word degree, Eq. (G8) bounds $\mathbb{E}\|\boldsymbol{M}\|_{\mathrm{op}}^2$ independently of n. The quadratic-form fluctuations therefore vanish in L^2, and the trace limits give the displayed Gram limit and source orthogonality.

Every finite controlled computation expands into finitely many mixed-word vectors with coefficients polynomial in the initial Gram entries and the scalar controls; normalized rank-one maps preserve this closure through $(p \otimes q)z = p\langle q, z\rangle$. Hence convergence is uniform on compact controls and transfers to tight realized controls along subsequences.

For coordinates, apply an independent Haar rotation to the entire computation. Simultaneous orthogonal conjugation preserves initialization, vector computations are equivariant, and scalar controls are invariant. Conditional on the unrotated computation, (ii) gives the Gaussian row limit determined by the normalized Gram matrix, uniformly on bounded Gram sets, including singular limits. Compact-uniform polynomial convergence replaces this Gram matrix by its continuous limiting formula evaluated at the realized controls. Subsequence convergence of the tight controls then gives the Gaussian mixture and conditional independence of distinct rows. For reference, (iii) gives

$$\mathbb{E}_{\boldsymbol{Q}}\left[|(\boldsymbol{Q}\boldsymbol{z})_i|^2\right] = \|\boldsymbol{z}\|_{\mathrm{rms}}^2, \qquad \mathbb{E}_{\boldsymbol{Q}}\left[|(\boldsymbol{Q}\boldsymbol{z})_i|^4\right] = \frac{3n}{n+2}\|\boldsymbol{z}\|_{\mathrm{rms}}^4. \tag{G10}$$

Initialization and control truncation events are rotation-invariant. These moment identities transfer Hilbert-norm bounds to coordinate tightness along arbitrary allowed position sequences; positive limiting variances and Eq. (F2) give coordinate anti-concentration, and increasing the control compacts removes truncation. □

Loopwise limit. At $(\alpha_N, \gamma_N) = (1, 0)$ and for fixed controls d_t, define the continuum recursion

$$\begin{aligned} C_{j,0} &= C_j, \\ R_{j,t} &= C_{j,t} \cdots C_{1,t}, \quad R_{0,t} = I, \\ S_{j,t} &= R_{j-1,t}, \quad T_{j,t} = C_{k,t} \cdots C_{j+1,t}, \\ B_t &= C_{k,t} \cdots C_{1,t} = R_{k,t}, \\ X_t(s) &= e^{asB_t} u_\star, \quad Y_t(s) = e^{a(1-s)B_t^*} v_\star, \\ z_{j,t}(s) &= S_{j,t} X_t(s), \quad b_{j,t}(s) = T_{j,t}^* Y_t(s), \\ J_{j,t} &= a \int_0^1 b_{j,t}(s) \otimes z_{j,t}(s) \, \mathrm{d}s, \\ C_{j,t+1} &= C_{j,t} - d_t J_{j,t}. \end{aligned} \tag{G11}$$

For fixed N, let $C_{j,t,N}, R_{j,t,N}, B_{t,N}, J_{j,t,N}$ denote the discrete counterpart obtained by replacing the exponentials by $E_{t,N}^s u_\star, (E_{t,N}^*)^{N-1-s} v_\star$, where $E_{t,N} = I + aB_{t,N}/N$, and the integral by the normalized sum over $s = 0, \ldots, N-1$.

Lemma 5. *The continuum recursion is well posed in operator and Hilbert–Schmidt norms, with all operators and paths bounded uniformly on compact control sets. Its paths are continuously differentiable in s and depend continuously on each finite control transcript. Moreover, uniformly on compact control sets as $N \to \infty$,*

$$\begin{aligned} \|C_{j,t,N} - C_{j,t}\|_{\mathrm{op}} + \|J_{j,t,N} - J_{j,t}\|_{\mathrm{HS}} &= \mathcal{O}(N^{-1}), \\ \sup_{r \le N} \left\| E_{t,N}^r u_\star - X_t(r/N) \right\| &= \mathcal{O}(N^{-1}), \\ \sup_{1 \le r \le N} \left\| R_{j,t,N} E_{t,N}^{r-1} u_\star - R_{j,t} X_t((r-1)/N) \right\| &= \mathcal{O}(N^{-1}). \end{aligned} \tag{G12}$$

Proof. The finite induction for Eq. (G5) applies in this fixed Hilbert space with $\|C_j\|_{\mathrm{op}} \le 2$, $\|u_\star\| = \|v_\star\| = 1$. Use $\|p \otimes q\|_{\mathrm{HS}} = \|p\|\|q\|$ to get the same bounds also for the Hilbert–Schmidt integrals and sums. For bounded D, F,

$$e^{asD} - e^{asF} = \int_0^{as} e^{(as-v)D}(D - F)e^{vF} \, \mathrm{d}v$$

gives a Lipschitz bound on each common operator ball. The one-step exponential Taylor bound and product telescoping give

$$\sup_{r \le N} \left\| (I + aD/N)^r - e^{arD/N} \right\|_{\mathrm{op}} \le C/N. \tag{G13}$$

For example, the single-step error is at most $a^2 \|D\|_{\mathrm{op}}^2 e^{a\|D\|_{\mathrm{op}}/N}/(2N^2)$. Products of k bounded operators are Lipschitz in those operators. Replacing the discrete paths in each J by exponentials thus has Hilbert–Schmidt error $\mathcal{O}(N^{-1})$, including the backward shift by one loop. The integrand in Eq. (G11) has uniformly bounded derivative in s, so its Riemann-sum error is also $\mathcal{O}(N^{-1})$. Its dependence on the operators is Lipschitz by the same estimates. Induction in the fixed training time proves Eq. (G12). It also proves continuity in controls and all asserted path bounds. At each fixed N, the discrete recursion is a finite controlled polynomial computation of precisely the type in Eq. (G9); normalized rank-one matrices become ordinary rank-one maps. This proves its width-limit interpretation. Then Eq. (G12) supplies the subsequent loop limit in one fixed Hilbert space. For actual tight controls first use the fixed-N uniform polynomial statement, and next the uniform compact-control estimates in Eq. (G12). Subsequence extraction supplies limiting controls if their laws are not unique. The moment identities in Eq. (G10) and the Gaussian row limit in Lemma 2 (ii) give the coordinate laws along each such subsequence. No continuity of the rule producing the controls is needed, and no interchange of the two limits is made. □

Lemma 6. *Retain the circular system of Definition 8, now with finite color space $H = \mathbb{R}^k$ and orthonormal basis $e_1, \ldots, e_k$, and write $C_j := C(e_j)$. The following entire word evaluation functional*

is used only to test whether a represented Hilbert vector is zero. For $A \geq 1$ set

$$\mathcal{E}_A = \left\{ h = \sum_{w,\rho} h_{w,\rho} e_{w,\rho} : \|h\|_A := \sum_{w,\rho} |h_{w,\rho}| A^{|w|} < \infty \right\},$$
$$\mathcal{E} = \bigcap_{A \geq 1} \mathcal{E}_A.$$

Here $e_{w,\rho}$ is the word basis of the source copy $\rho \in \{u, v\}$ of the full Fock space in Definition 8. Complexify these spaces when evaluating at a complex number. Every vector path in Eq. (G11)*, and every finite operator product applied to one of these paths, belongs to $\mathcal{E}$. For each complex ω there is a linear functional Φ_ω on $\mathcal{E}$, continuous on every $\mathcal{E}_A$ with $A \geq 1 + |\omega|$, such that*

$$\begin{aligned} \Phi_\omega(u_\star) &= 1, \\ \Phi_\omega(v_\star) &= 0, \\ \Phi_\omega(C_j h) &= \omega \Phi_\omega(h), \\ \Phi_\omega(C_j^* h) &= 0. \end{aligned} \tag{G14}$$

It commutes with the path derivatives and integrals used below. It need not be bounded in Hilbert norm.

Proof. Write $e_{j,+} = (e_j, 0)$ and $e_{j,-} = (0, e_j)$, and for $b = (j, \pm)$ let $\bar{b} = (j, \mp)$ and $D_b = \mathsf{L}(e_b) + \mathsf{L}(e_{\bar{b}})^*$. Put $x_{(j,+)} = \omega$ and $x_{(j,-)} = 0$, and define on word bases

$$p_\varnothing(\omega) = 1, \qquad p_{bw}(\omega) = x_b p_w(\omega) - \mathbb{1}_{w = \bar{b} w'}\, p_{w'}(\omega). \tag{G15}$$

For $A = 1 + |\omega|$, the inequality $|\omega|/A + A^{-2} \leq 1$ proves inductively that $|p_w(\omega)| \leq A^{|w|}$. Thus $\Phi_\omega(h) = \sum_w h_{w,u} p_w(\omega)$ converges absolutely with $|\Phi_\omega(h)| \leq \|h\|_A$ whenever $A \geq 1 + |\omega|$. The identity $D_b e_w = e_{bw} + \mathbb{1}_{w = \bar{b} w'}\, e_{w'}$ proves the intertwining relations on finite words. Since

$$\|\mathsf{L}(e_b) h\|_A = A \|h\|_A, \qquad \|\mathsf{L}(e_b)^* h\|_A \leq A^{-1} \|h\|_A,$$

creation and annihilation are bounded on $\mathcal{E}_A$, and continuity extends those relations from finite words to $\mathcal{E}_A$.

Fix $A \geq 1$. Its weighted norm dominates the Hilbert norm, and both vacua belong to $\mathcal{E}_A$. Suppose the operators $C_{j,t}$ and their Hilbert adjoints act boundedly on $\mathcal{E}_A$, as they do at $t = 0$. Their finite products and norm-convergent exponential series are bounded there, giving depth-continuous paths $X_t, Y_t, z_{j,t}, b_{j,t}$ in $\mathcal{E}_A$. For real $p, q \in \mathcal{E}_A$,

$$\begin{aligned} \|(p \otimes q) h\|_A &\leq \|p\|_A \|q\|_{\mathcal{H}} \|h\|_A, \\ \|(p \otimes q)^* h\|_A &\leq \|q\|_A \|p\|_{\mathcal{H}} \|h\|_A, \qquad \mathcal{H} = \mathcal{H}_u \oplus \mathcal{H}_v. \end{aligned}$$

Consequently the rank-one integrands defining $J_{j,t}$ and $J_{j,t}^*$ are continuous bounded-operator-valued functions on the compact depth interval. Their Bochner integrals are bounded on $\mathcal{E}_A$, so the operator updates and their adjoints satisfy the same induction at $t + 1$.

This holds for every fixed A, proving that all stated paths and finite operator products belong to $\mathcal{E}$. The exponential partial sums agree in the weighted and Hilbert spaces, hence so do their limits and the update integrals. Bounded-operator exponentials make the paths differentiable in the loop position and continuous in the finite control transcript, with bounds on compact control sets for each fixed A. For a chosen ω, continuity of Φ_ω in one $\mathcal{E}_A$ with $A \geq 1 + |\omega|$ therefore permits the depth derivatives and integrals used below. In complex expressions the real inner-product functionals are extended complex-linearly. Nonvanishing is tested by this evaluator; Hilbert-norm lower bounds subsequently follow from continuity and compactness. □

Lemma 7. *For every finite real control transcript in Eq.* (G11)*, every $0 \leq j \leq k$, and every $0 \leq s \leq 1$,*

$$R_{j,t} X_t(s) \neq 0. \tag{G16}$$

For each compact control set these norms have positive lower bounds uniform in s and in the finitely many j, t under consideration. In particular this applies to $B_t e^{a s B_t} u_\star = R_{k,t} X_t(s)$.

Proof. We prove a stronger asymptotic identity using Eq. (G14). Let $r \geq 1$ and choose

$$\omega = re^{\mathrm{i}\pi/(2k)}, \qquad \omega^k = \mathrm{i}r^k.$$

All estimates in this proof are uniform over $s \in [0,1]$ and over a fixed compact set of finite real controls. Constants may depend on that compact set, a, k, t, but not on r. Write $F_{j,t}(s,\omega) = \Phi_\omega(R_{j,t}X_t(s))$. We claim

$$F_{j,t}(s,\omega) = \omega^j e^{as\omega^k} + \mathcal{O}(r^{j-k}), \qquad 0 \leq j \leq k. \tag{G17}$$

At $t = 0$, Eq. (G14) and the exponential power series give exact equality without the error. We now establish the statement by induction over the training time, including the bounds on the error functionals. Put

$$K_{m,t} = C_{m,t} - C_m = -a\sum_{q<t} d_q \int_0^1 b_{m,q}(s) \otimes R_{m-1,q}X_q(s)\ \mathrm{d}s. \tag{G18}$$

Suppose Eq. (G17) is known at every $q < t$. For $h \in \mathcal{E}$, applying Φ_ω to the adjoint of Eq. (G18) and using $\Phi_\omega(C_m^* h) = 0$ gives

$$\Phi_\omega(C_{m,t}^* h) = -a\sum_{q<t} d_q \int_0^1 F_{m-1,q}(s,\omega)\langle b_{m,q}(s), h\rangle\ \mathrm{d}s. \tag{G19}$$

The scalar function $g(s) = \langle b_{m,q}(s), h\rangle$ has $\sup_s(|g(s)|+|g'(s)|) \leq C\|h\|$, by Eq. (G11). Integration by parts therefore gives

$$\left|\int_0^1 e^{\mathrm{i}ar^k s} g(s)\ \mathrm{d}s\right| \leq \frac{|g(0)| + |g(1)| + \int_0^1 |g'(s)|\ \mathrm{d}s}{ar^k} \leq Cr^{-k}\|h\|.$$

Inserting Eq. (G17) for $F_{m-1,q}$ in Eq. (G19), the leading term contributes $\mathcal{O}(r^{m-1-k})\|h\|$; its error has the same bound by direct integration. Hence

$$|\Phi_\omega(C_{m,t}^* h)| \leq Cr^{m-1-k}\|h\|. \tag{G20}$$

This composite functional, unlike Φ_ω itself, extends to a bounded Hilbert-space functional with the displayed norm. We need corresponding estimates for the backward range vectors. For $j < k$,

$$b_{j,t}(s) = C_{j+1,t}^* \left(C_{j+2,t}^* \cdots C_{k,t}^* Y_t(s)\right).$$

Use Eq. (G20) with $m = j+1$, and the Hilbert bounds from Eq. (G11), to obtain

$$|\Phi_\omega(b_{j,t}(s))| \leq Cr^{j-k} \quad (j < k). \tag{G21}$$

For $j = k$, a slightly sharper bound is useful. Since $B_t^* = C_{1,t}^* \cdots C_{k,t}^*$, Eq. (G20) with $m = 1$ shows that $\Phi_\omega B_t^*$ is bounded in Hilbert norm by Cr^{-k}. As $\Phi_\omega(v_\star) = 0$,

$$\Phi_\omega(e^{avB_t^*} v_\star) = a\int_0^v \Phi_\omega(B_t^* e^{auB_t^*} v_\star)\ \mathrm{d}u = \mathcal{O}(r^{-k}), \qquad 0 \leq v \leq 1.$$

Thus

$$|\Phi_\omega(b_{k,t}(s))| \leq Cr^{-k}. \tag{G22}$$

For time zero Eqs. (G21) and (G22) also hold: all these evaluations are zero. We retain these two estimates as additional induction conclusions. Next apply Φ_ω to Eq. (G18) itself. The bounds in Eqs. (G21) and (G22) at the earlier times $q < t$ and the bounded forward paths show

$$|\Phi_\omega(K_{m,t}h)| \leq \begin{cases} Cr^{m-k}\|h\|, & m < k, \\ Cr^{-k}\|h\|, & m = k. \end{cases} \tag{G23}$$

In particular each of these functionals extends boundedly to $\mathcal{H}_u \oplus \mathcal{H}_v$. Use $\Phi_\omega(C_m h) = \omega\Phi_\omega(h)$ successively along a prefix. We obtain the exact identity on $\mathcal{E}$

$$\begin{aligned} \Phi_\omega(R_{j,t}h) &= \omega^j \Phi_\omega(h) + D_{j,t,\omega}(h), \\ D_{j,t,\omega}(h) &= \sum_{m=1}^{j} \omega^{j-m}\Phi_\omega(K_{m,t}R_{m-1,t}h). \end{aligned} \tag{G24}$$

The initial term $D_{0,t,\omega}$ is zero. By Eq. (G23) and the Hilbert operator bounds for the prefixes, $D_{j,t,\omega}$ extends to a bounded Hilbert functional satisfying

$$\|D_{j,t,\omega}\|_{(\mathcal{H}_u \oplus \mathcal{H}_v)^*} \le C r^{j-k} \quad (1 \le j \le k). \tag{G25}$$

For $j = k$ the term $m = k$ has the stronger bound Cr^{-k}; the other terms are $\mathcal{O}(1)$. For $k = 1$ there are no other terms, and the asserted $\mathcal{O}(1)$ bound is still valid. Now $X_t'(s) = aB_t X_t(s)$, so Eq. (G24) for $j = k$ gives the scalar ODE

$$\begin{aligned} F_{0,t}'(s,\omega) &= a\omega^k F_{0,t}(s,\omega) + aH_{t,\omega}(s), \\ F_{0,t}(0,\omega) &= 1, \\ H_{t,\omega}(s) &= D_{k,t,\omega}(X_t(s)). \end{aligned}$$

The crucial consequence of Eq. (G25) is

$$\sup_s \big(|H_{t,\omega}(s)| + |H_{t,\omega}'(s)|\big) \le C, \tag{G26}$$

independently of r: both X_t and $X_t' = aB_t X_t$ have bounded Hilbert norms. Solving the ODE gives

$$F_{0,t}(s,\omega) = e^{as\omega^k} + a \int_0^s e^{a(s-v)\omega^k} H_{t,\omega}(v)\, \mathrm{d}v.$$

Another integration by parts, with Eq. (G26) and $|e^{a(s-v)\omega^k}| = 1$, bounds this last integral by Cr^{-k}, uniformly even at $s = 0$. This proves Eq. (G17) for $j = 0$. Applying Eq. (G24) with $h = X_t(s)$ and using Eq. (G25) now proves it for all j: multiplication of the $j = 0$ error by ω^j has size $\mathcal{O}(r^{j-k})$. This completes the induction. The leading term in Eq. (G17) has modulus exactly r^j. Choose r sufficiently large that its error is at most $r^j/2$. Then $F_{j,t}(s,\omega) \neq 0$ at every s, proving Eq. (G16), because the zero vector has zero evaluation on the stated domain. This uses no sign or smallness restriction on the controls. We do not infer a Hilbert lower bound from the unbounded evaluator. Instead, Eq. (G11) shows that $(d_0, \dots, d_{t-1}, s) \mapsto \|R_{j,t} X_t(s)\|$ is continuous in Hilbert norm. It is everywhere positive by Eq. (G16), so compactness gives its positive minimum on a compact control set times $[0, 1]$. There are finitely many j, t, so their joint minimum is positive as well. □

First-step response.

Proposition 8. *At $(\alpha_N, \gamma_N) = (1, 0)$, for every fixed $a, \eta > 0$, the first update with a nondegenerate initial training signal has a nonzero ordered-limit output response. For every fixed $0 < \lambda \le 1$, its changes in $\boldsymbol{x}$ and in all $\boldsymbol{h}^{r,j}$ at $r = \lfloor \lambda N \rfloor$ have nonzero $\Theta(1)$ coordinatewise scale on the declared input batch.*

Proof. Use $d_\star = \eta \chi_\star \xi_0$ and write $B = C_k \cdots C_1$, $S_j = C_{j-1} \cdots C_1$, $T_j = C_k \cdots C_{j+1}$, $R_j = C_j \cdots C_1$, and let $J_j = J_{j,0}$ from Eq. (G11). The initial matrices and all their products preserve $\mathcal{H}_u$ and $\mathcal{H}_v$. Every J_j maps $\mathcal{H}_u$ into $\mathcal{H}_v$ and vanishes on $\mathcal{H}_v$, because its range vectors belong to the latter and its domain vectors to the former. Thus every product containing two J's, separated by any products of initial matrices, is zero. The same is true for the discrete $J_{j,0,N}$. It follows, for the simultaneous update of all matrices, that

$$\begin{aligned} B_1 &= B - d_\star H, \\ H &= \sum_{j=1}^{k} T_j J_j S_j, \\ R_{j,1} &= R_j - d_\star M_j, \\ M_j &= \sum_{m=1}^{j} C_j \cdots C_{m+1} J_m S_m. \end{aligned} \tag{G27}$$

Here H, M_j also map the first source into the second and vanish on the second. In particular all powers and exponentials involving these perturbations contain at most one perturbation. Duhamel yields the exact identity

$$\begin{aligned} X_1(\lambda) - X_0(\lambda) &= -d_\star D(\lambda), \\ D(\lambda) &= a \int_0^\lambda e^{a(\lambda - s)B} H e^{asB} u_\star \, \mathrm{d}s. \end{aligned} \tag{G28}$$

There is no Taylor remainder in $d_\star$. Pair with $v_\star$ at $\lambda = 1$. The contribution of matrix j is

$$a \int_0^1 \langle T_j^* e^{a(1-s)B^*} v_\star, J_j S_j e^{asB} u_\star \rangle \, \mathrm{d}s = \langle J_j, J_j \rangle_{\mathrm{HS}}. \tag{G29}$$

The equality follows by inserting the defining rank-one integral for one of the two J_j's. Therefore

$$f_1(\zeta) - f_0(\zeta) \xrightarrow{\mathbb{P}} -d_\star \zeta \sum_{j=1}^{k} \|J_j\|_{\mathrm{HS}}^2. \tag{G30}$$

At fixed N the identical computation with residual powers and $J_{j,0,N}$ gives the same formula with those discrete J's. Thus Eq. (G12) proves the actual ordered limit, including all simultaneous matrix-update terms. For positivity, expand the two exponentials in the integral defining J_j. Their constant terms give the coefficient a of the unit Hilbert–Schmidt basis vector $(T_j^* v_\star) \otimes (S_j u_\star)$, while every remaining all-creation term is orthogonal to it. Hence, directly within this proposition,

$$\|J_j\|_{\mathrm{HS}} \geq a.$$

Hence Eq. (G30) is nonzero for every nondegenerate input. At $\zeta = \xi_0$ it equals $-\eta \chi_\star \xi_0^2 \sum_j \|J_j\|_{\mathrm{HS}}^2$. We give an explicit coordinate projection to prove all feature lower bounds. All coefficients of every initial C_j, C_j^*, their products, and their exponentials at nonnegative times are nonnegative in the word basis. The same is true of J_j, H, M_j and $D(\lambda)$. The $j = k$ term of H is $J_k S_k$. For $0 \leq s \leq 1$,

$$\begin{aligned} \langle v_\star, J_k S_k e^{asB} u_\star \rangle &= a \int_0^1 \langle v_\star, e^{a(1-q)B^*} v_\star \rangle \langle S_k e^{aqB} u_\star, S_k e^{asB} u_\star \rangle \, \mathrm{d}q \\ &\geq a. \end{aligned} \tag{G31}$$

Indeed the first factor equals one and the second is at least one: the same-direction prefix vectors have a common shortest word with coefficient one. The identity term in the outer exponential in Eq. (G28) therefore gives

$$\langle v_\star, D(\lambda) \rangle \geq a^2 \lambda. \tag{G32}$$

For the internal response use Eqs. (G27) and (G28). Terms with two perturbations vanish, so exactly

$$R_{j,1} X_1(\lambda) - R_j X_0(\lambda) = -d_\star \left(M_j X_0(\lambda) + R_j D(\lambda) \right). \tag{G33}$$

The two vectors inside parentheses have nonnegative coefficients. Moreover $R_j v_\star = e_{(j,+)\cdots(1,+),v}$. Retaining the vacuum coefficient from Eq. (G32) in $R_j D(\lambda)$ shows that the norm of the vector in parentheses is at least $a^2\lambda$. Thus both Eqs. (G28) and (G33) have norm at least $|d_\star| a^2 \lambda > 0$, with finite upper bounds from bounded operators. The lower constant may depend on the fixed positive λ; no lower bound at $\lambda = 0$ is required for feature learning. The convergence in Eq. (G12) transfers these vectors and their Gram norms to the ordered finite-network limit; the internal feature's depth $(\lfloor \lambda N \rfloor - 1)/N$ differs from λ by $\mathcal{O}(N^{-1})$. The coordinate kernel in Lemma 4 gives nondegenerate centered Gaussian laws with these squared norms as variances. Scalar nondegenerate inputs and fixed finite batches preserve the stated scales. This proves feature learning at the update $t = 0 \to 1$. □

For $k = 1$, set $B_t = C_{1,t}$ and $G_t = J_{1,t}$. Then Eq. (G11) becomes

$$B_0 = C, \qquad G_t = a \int_0^1 e^{a(1-s)B_t^*} v_\star \otimes e^{asB_t} u_\star \, \mathrm{d}s, \qquad B_{t+1} = B_t - d_t G_t.$$

Lemma 7 with $j = 1$ gives $B_t e^{asB_t} u_\star \neq 0$ throughout $[0, 1]$. At the first update, Proposition 8 gives output response $-d_\star \zeta \|G_0\|_{\mathrm{HS}}^2$ and state projection of magnitude at least $|d_\star| a^2 \lambda$. These specializations also apply to a one-color invariant subspace of a larger circular system.

G.2 Classification of loopwise parameterizations at fixed block depth

The following results apply to every fixed finite $k \geq 1$; the fixed-k Claims use them for the $k \geq 2$ extension, while the additional $k = 1$ classification follows in Appendix G.3. The initialization threshold and gradient normalization follow from Lemma 9; Lemma 10 supplies necessity for training stability, and the finite-step bounds, response, and prefix lower bounds prove Claims $2'$ and $3'$. Nontriviality is classified only within the initialization- and training-stable region.

Lemma 9 (Initialization and matrix-gradient normalization)**.** *Initialization is stable if and only if $\alpha_N \geq 1$, and is faithful there on every declared nondegenerate batch. For $\alpha_N \geq 1$ and $\gamma_N \geq 0$, every fixed SGD routine satisfies, for every fixed matrix j and training step t,*

$$\boldsymbol{G}_{j,t} = \mathcal{O}\left(n^{-1}N^{1-\alpha_N}\right). \tag{G34}$$

For a nondegenerate initial training signal at $t = 0$, this scale is attained at every fixed entry, including diagonal entries. Consequently $\widetilde{\boldsymbol{G}}_{j,t} = nN^{\alpha_N-1}\boldsymbol{G}_{j,t}$ is the faithful shared-block gradient normalization for every matrix j.

Proof. For $\alpha_N \geq 1$, Lemma 3 gives bounded normalized states and outputs at initialization, and Eq. (G10) gives the coordinatewise upper bounds uniformly over loop position. For necessity, write $B = C_k \cdots C_1$ and $w = (k, +) \cdots (1, +)$. Every multiplication in $B^m u_\star$ is a creation, so $B^m u_\star = e_{w^m,u}$, and these vectors are orthonormal. Consequently the fixed-N limiting squared norm of $\boldsymbol{A}_0^N \boldsymbol{u}$ is

$$\sigma_N^2 = \sum_{m=0}^{N} \binom{N}{m}^2 c_N^{2m} \geq 1 + a^2 N^{2-2\alpha_N}. \tag{G35}$$

By Lemma 4, a fixed coordinate converges at fixed N to $\mathcal{N}(0, \sigma_N^2)$. If $\alpha_N < 1$, then for every $M > 0$ its subsequent probability of lying in $[-M, M]$ is at most $\sqrt{2/\pi}M/\sigma_N \to 0$, so initialization stability fails.

For initialization faithfulness, $R_j B^m u_\star$ are also orthonormal as m varies. In loop r, the input-one preactivation variance is therefore $\sum_{m=0}^{r-1} \binom{r-1}{m}^2 c_N^{2m}$, between 1 and e^{2a} when $\alpha_N \geq 1$. The Gaussian coordinate law gives the claimed lower scale uniformly over loops and the fixed set of matrices.

For the upper gradient bound, work first on an invariant compact-control truncation event. Every $\boldsymbol{b}, \boldsymbol{z}$ in Eq. (G5) then has normalized norm at most a fixed C. For fixed entry indices i, l, Eq. (G10) and Cauchy–Schwarz give

$$\mathbb{E}_{\boldsymbol{Q}} \left|(\boldsymbol{Q}\boldsymbol{b})_i (\boldsymbol{Q}\boldsymbol{z})_l\right|^2 \leq 3C^4.$$

The $L^2(\boldsymbol{Q})$ triangle inequality bounds the sum over N loops by $\sqrt{3}C^2 N$, without requiring independence between loops. Apply Eq. (G5), use compact bounds on $\chi_t \xi_t$, and then remove truncation to obtain Eq. (G34).

For the matching first-step scale, retain $\xi_0 \neq 0$ and $\chi_0 \to \chi_\star \neq 0$, and put $S_j = C_{j-1} \cdots C_1$ and $T_j = C_k \cdots C_{j+1}$ at initialization. The families

$$\begin{aligned} f_{j,p} &= S_j B^p u_\star, \\ e_{j,q} &= T_j^* (B^*)^q v_\star, \qquad p, q \geq 0, \end{aligned}$$

are orthonormal within each family and orthogonal across the two source copies. Expanding the exact gradient and using $\sum_{s=0}^{N-1} \binom{s}{p}\binom{N-1-s}{q} = \binom{N}{p+q+1}$ gives the coefficient of $e_{j,q} \otimes f_{j,p}$ in

$$J_{j,N} \coloneqq \frac{a}{N} \sum_{s<N} T_j^* (I + c_N B^*)^{N-1-s} v_\star \otimes S_j (I + c_N B)^s u_\star$$

as

$$b_{qp,N} = \frac{a}{N} c_N^{p+q} \binom{N}{p+q+1}. \tag{G36}$$

Take this coefficient to be zero when $p + q + 1 > N$. For $\alpha_N \geq 1$,

$$0 \leq b_{qp,N} \leq \frac{a^{p+q+1}}{(p+q+1)!}, \qquad b_{00,N} = a.$$

The coefficient matrices converge in Hilbert–Schmidt norm to

$$b_{qp} = \begin{cases} a^{p+q+1}/(p+q+1)!, & \alpha_N = 1, \\ a\, \mathbb{1}_{p=q=0}, & \alpha_N > 1. \end{cases} \tag{G37}$$

At fixed N, each fixed width-limit entry of $nN^{\alpha_N-1}\boldsymbol{G}_{j,0}$ has law $\chi_\star\xi_0\sum_{p,q} b_{qp,N}E_qF_p$ for independent standard Gaussian sequences E, F. The coupling from Eq. (G37) gives L^2 convergence to $\chi_\star\xi_0\sum b_{qp}E_qF_p$. Conditional on F, this is centered Gaussian with finite and almost surely positive variance $(\chi_\star\xi_0)^2\|bF\|_{\ell^2}^2$. Hence the limiting entry is nonzero almost surely and tight, which proves the matching weak scale claimed in the lemma. □

Lemma 10. *Assume $\alpha_N \geq 1$ and $\gamma_N < 0$. For a first update with a nondegenerate initial training signal, training stability fails: already $\Delta\boldsymbol{h}_0^{1,1}$ is not $\mathcal{O}(1)$ in the ordered coordinatewise limit.*

Proof. Here $d_0 = \eta\chi_0\xi_0 \to d_\star \neq 0$ in the width limit. At finite width, set

$$\boldsymbol{J}_{1,N}^{(n)} \coloneqq \frac{a}{N}\sum_{s=0}^{N-1}\frac{\boldsymbol{b}_{1,0,s}\boldsymbol{z}_{1,0,s}^{\mathsf{T}}}{n}.$$

With $B = C_k\cdots C_1$, $S_1 = I$, $T_1 = C_k\cdots C_2$, and $E_N = I + aN^{-\alpha_N}B$, its fixed-$N$ width-limit representative is

$$J_{1,N} = \frac{a}{N}\sum_{s=0}^{N-1} T_1^*(E_N^*)^{N-1-s}v_\star \otimes S_1E_N^s u_\star.$$

The exact update in Eq. (G5) gives

$$\boldsymbol{W}_1^1 - \boldsymbol{W}_0^1 = -d_0N^{-\gamma_N}\boldsymbol{J}_{1,N}^{(n)}.$$

By Eq. (G36), the coefficient of $T_1^*v_\star \otimes u_\star$ in $J_{1,N}$ is $b_{00,N} = a$. All other all-creation coefficients are orthogonal, and therefore

$$\|J_{1,N}u_\star\| \geq a. \tag{G38}$$

On the fixed input $\zeta = 1$, since $\boldsymbol{h}_0^{1,1} = \boldsymbol{W}_0^1\boldsymbol{u}$ and $\boldsymbol{h}_1^{1,1} = \boldsymbol{W}_1^1\boldsymbol{u}$, the finite-width update is

$$\Delta\boldsymbol{h}_0^{1,1} = -d_0N^{-\gamma_N}\boldsymbol{J}_{1,N}^{(n)}\boldsymbol{u}.$$

By Lemma 4, the normalized Gram norm of $\boldsymbol{J}_{1,N}^{(n)}\boldsymbol{u}$ converges, with N fixed, to $\|J_{1,N}u_\star\|$, and each fixed coordinate converges to the corresponding centered Gaussian limit. The random-rotation moment identities in Eq. (G10) and Eq. (G38) therefore imply a uniform Paley–Zygmund bound: there exist $c, p > 0$, independent of N, such that

$$\mathbb{P}\{|Z_{N,i}| \geq c\} \geq p,$$

where $Z_{N,i}$ denotes the fixed-coordinate width-limit variable represented by $J_{1,N}u_\star$. Because $|d_\star|N^{-\gamma_N} \to \infty$, the coordinates of $\Delta\boldsymbol{h}_0^{1,1}$ are not tight in the subsequent $N \to \infty$ limit. Thus training stability fails for every $\gamma_N < 0$, independently of the fixed finite k. □

Proof of Claim 2′. Within the initialization-stable and training-stable region, we prove that nontriviality holds if and only if

$$\alpha_N + \gamma_N \leq 1. \tag{G39}$$

Initialization stability is equivalent to $\alpha_N \geq 1$ by Lemma 9. If $\gamma_N < 0$, Lemma 10 supplies a nondegenerate first step whose internal update is not $\mathcal{O}(1)$. If $\alpha_N \geq 1$ and $\gamma_N \geq 0$, Eqs. (G6) and (G10) give coordinatewise tightness of every state change and the output change, uniformly over loop position; compact truncation covers arbitrary fixed SGD routines. This proves the training-stability equivalence.

Summing Eq. (G6) over the fixed number of updates gives

$$|f_t(\zeta) - f_0(\zeta)| = \mathcal{O}\left(N^{1-\alpha_N-\gamma_N}\right)$$

on every fixed bounded input batch. Thus $\alpha_N + \gamma_N > 1$ implies triviality. Within the stable region, $\alpha_N + \gamma_N \leq 1$ forces $(\alpha_N, \gamma_N) = (1, 0)$, and Eq. (G30) proves nontriviality there. This establishes Eq. (G39) and uniqueness. □

Proof of Claim 3′. Claim 2′ proves that $(1, 0)$ is the unique stable and nontrivial point. We now prove its feature and diversity properties.

At the critical point, Lemma 7 and compactness give constants $0 < m \le M < \infty$, uniform in the loop fraction and in the finitely many matrix indices, such that each of $\|X_t(s)\|$ and $\|R_{j,t}X_t(s)\|$ lies in $[m, M]$ on every compact set of finite control transcripts. By Eqs. (G9) and (G12), these are the ordered normalized Gram norms of the finite-network residual and internal features. Their fixed coordinates are the corresponding Gaussian mixtures. Let $\mathcal{D}_t = (d_0, \dots, d_{t-1})$ be the limiting scalar control vector and $\mathcal{D}_t = \sigma(\mathcal{D}_t)$. On a fixed compact-control truncation, every such coordinate satisfies

$$Z \mid \mathcal{D}_t \sim \mathcal{N}\left(0, \sigma_Z^2(\mathcal{D}_t)\right), \qquad \sigma_Z(\mathcal{D}_t) \ge c_\varsigma m \quad \text{a.s.}$$

Consequently, for every $r > 0$,

$$\mathbb{P}\left(|Z| \le r \mid \mathcal{D}_t\right) \le \sqrt{\frac{2}{\pi}}\,\frac{r}{c_\varsigma m} \quad \text{a.s.} \tag{G40}$$

Together with the uniform upper-tail bounds, truncation proves the weak $\Theta(1)$ scale for all routines. Equivalently, for each fixed routine and training step,

$$\boldsymbol{H}_t^{r,j}(\boldsymbol{\zeta}) = \Theta(1), \qquad \boldsymbol{X}_t^{r}(\boldsymbol{\zeta}) = \Theta(1),$$

uniformly over $1 \le r \le N$ and the fixed set $1 \le j \le k$. The second assertion is the strengthened residual-state conclusion. First-step feature learning follows from Proposition 8.

Finally, Taylor expansion in the bounded limiting operator gives

$$\begin{aligned} X_t(\lambda + \varepsilon) - X_t(\lambda) &= a\varepsilon B_t X_t(\lambda) + E_t(\lambda, \varepsilon), \\ \|E_t(\lambda, \varepsilon)\| &\le C\varepsilon^2. \end{aligned} \tag{G41}$$

The bound is uniform on compact controls and $0 \le \lambda \le 1 - \varepsilon$. By Lemma 7, the leading-vector norm has a positive minimum, so the norm in Eq. (G41) is $\Theta(\varepsilon)$. The limits in Eqs. (G9) and (G12) transfer this to the normalized finite-network batch norm; floor errors are $\mathcal{O}(N^{-1})$ and vanish before $\varepsilon \downarrow 0$. Hence $\kappa_N = 0$, and the matching upper bound at initialization excludes every $\kappa_N > 0$. □

G.3 Sharper nontriviality classification at block depth 1

For $k = 1$ write $\boldsymbol{W}_t = \boldsymbol{W}_t^1$, $\boldsymbol{z}_{t,s} = \boldsymbol{z}_{1,t,s}$, and $\boldsymbol{b}_{t,s} = \boldsymbol{b}_{1,t,s}$. The rescaled parameter recursion below controls the output even when $\gamma_N < 0$, and the exact first-step response supplies a matching lower bound. This extends the nontriviality classification to the whole initialization-stable region.

Lemma 11. *Suppose $\alpha_N \ge 1$, and allow any $\gamma_N \in \mathbb{R}$. Put*

$$\beta_N = N^{1-\alpha_N}, \qquad \rho_N = N^{1-\alpha_N-\gamma_N}. \tag{G42}$$

If $\alpha_N + \gamma_N > 1$, every fixed SGD routine is trivial:

$$f_t(\zeta) - f_0(\zeta) = \mathcal{O}(\rho_N) \to 0. \tag{G43}$$

For a nondegenerate initial training signal, at every fixed N, the width-limit first output response is exactly

$$f_1(\zeta) - f_0(\zeta) = -\eta\chi_\star\xi_0\zeta\,\rho_N\|J_N\|_{\mathrm{HS}}^2, \tag{G44}$$

where

$$\begin{aligned} E_N &= I + \frac{a\beta_N}{N}C, \\ J_N &= \frac{a}{N}\sum_{s=0}^{N-1}(E_N^*)^{N-1-s}v_\star \otimes E_N^s u_\star. \end{aligned} \tag{G45}$$

Uniformly over N,

$$\begin{aligned} a^2 \le \|J_N\|_{\mathrm{HS}}^2 \le a^2 e^{4a}, \\ a \le \|J_N u_\star\| \le a e^{2a}. \end{aligned} \tag{G46}$$

At $\zeta = \xi_0$ the coefficient is $-\eta\chi_\star\xi_0^2$. Consequently the model is nontrivial whenever $\alpha_N + \gamma_N \le 1$.

Proof. We first prove the universal upper assertion, including points with $\gamma_N < 0$, for which the unscaled matrix norm need not be bounded. Define $\widetilde{\boldsymbol{W}}_t = \beta_N \boldsymbol{W}_t$. Then

$$\begin{aligned} \boldsymbol{A}_t &= \boldsymbol{I} + \frac{a}{N}\widetilde{\boldsymbol{W}}_t, \\ \widetilde{\boldsymbol{W}}_{t+1} - \widetilde{\boldsymbol{W}}_t &= -\frac{a d_t \rho_N}{N} \sum_{s=0}^{N-1} \frac{\boldsymbol{b}_{t,s}\boldsymbol{z}_{t,s}^{\mathsf{T}}}{n}. \end{aligned} \tag{G47}$$

Here $\boldsymbol{b}, \boldsymbol{z}, d$ are the exact finite-network quantities in Eq. (G5). When $\alpha_N + \gamma_N \geq 1$, both β_N and ρ_N are at most one. On the invariant initialization and compact-control events used in Lemma 3, $\|\widetilde{\boldsymbol{W}}_0\|_{\mathrm{op}} \leq \|\boldsymbol{W}_0\|_{\mathrm{op}}$. If $\|\widetilde{\boldsymbol{W}}_t\|_{\mathrm{op}} \leq R_t$, then

$$\begin{aligned} \|\boldsymbol{z}_{t,s}\|_{\mathrm{rms}}, \|\boldsymbol{b}_{t,s}\|_{\mathrm{rms}} &\leq B_0 e^{aR_t}, \\ \left\|\widetilde{\boldsymbol{W}}_{t+1} - \widetilde{\boldsymbol{W}}_t\right\|_{\mathrm{op}} &\leq a|d_t|\rho_N B_0^2 e^{2aR_t}. \end{aligned}$$

Thus the same finite induction as in Lemma 3 bounds all $\|\widetilde{\boldsymbol{W}}_t\|_{\mathrm{op}}$ independently of n, N, with the update norm $\mathcal{O}(\rho_N)$. Telescoping the N residual maps gives

$$\left\|(\boldsymbol{A}_{t+1}^r - \boldsymbol{A}_t^r)\boldsymbol{u}\right\|_{\mathrm{rms}} \leq C\rho_N \quad (r \leq N).$$

Pair with $\boldsymbol{v}/n$, multiply by the bounded input, and sum over finitely many training steps to prove Eq. (G43). Compact truncation handles the realized tight controls. This is a bound on the actual finite networks, so it respects the specified order of limits. It does not assert that the unscaled preactivations are bounded. For the lower bounds, fix N first and let $d_\star = \eta\chi_\star\xi_0$. Lemma 4 applies to any fixed finite coefficient, including $N^{-\gamma_N}$. Thus the width-limit initial operator is C, its residual map is E_N, and its first update is exactly

$$B_{1,N} = C - d_\star N^{-\gamma_N} J_N.$$

The two source summands are orthogonal and preserved by C, E_N. Every range vector in J_N belongs to $\mathcal{H}_v$, whereas every domain vector belongs to $\mathcal{H}_u$. Hence $J_N E_N^k J_N = 0$ for every $k \geq 0$. Write $c_N = a\beta_N/N$. Expanding the residual power therefore yields the exact identity

$$(E_N - d_\star c_N N^{-\gamma_N} J_N)^N - E_N^N = -d_\star c_N N^{-\gamma_N} \sum_{q=0}^{N-1} E_N^{N-1-q} J_N E_N^q. \tag{G48}$$

There are no higher-order terms in the update. Pairing with the two Fock vacua, the sum on the right is

$$\sum_{q=0}^{N-1} \langle (E_N^*)^{N-1-q} v_\star, J_N E_N^q u_\star \rangle = \frac{N}{a}\|J_N\|_{\mathrm{HS}}^2.$$

The initial scalar contraction is zero. Since $c_N N^{-\gamma_N}(N/a) = \rho_N$, this proves Eq. (G44), as an equality of fixed-N width limits. For completeness, the bounds in Eq. (G46) require no continuum approximation. The vectors $E_N^s u_\star$ have nonnegative coefficients along $e_{(1,+)^m,u}$, with vacuum coefficient one; the backward vectors have the corresponding property along $e_{(1,-)^m,v}$. All pairwise inner products within either family are at least one. Expanding the squared Hilbert–Schmidt norm in Eq. (G45) therefore gives

$$\|J_N\|_{\mathrm{HS}}^2 = \frac{a^2}{N^2} \sum_{s,q=0}^{N-1} \langle (E_N^*)^{N-1-s} v_\star, (E_N^*)^{N-1-q} v_\star \rangle \langle E_N^s u_\star, E_N^q u_\star \rangle \geq a^2.$$

Also each vector in these two same-direction families has norm at most $(1 + c_N)^N \leq e^a$, giving the asserted upper bound. Since $\langle E_N^s u_\star, u_\star \rangle = 1$ and every backward vacuum coefficient is one, $\langle v_\star, J_N u_\star \rangle = a$. The upper vector bound follows from $\|J_N\|_{\mathrm{op}} \leq \|J_N\|_{\mathrm{HS}}$. This proves Eq. (G46). If $\alpha_N + \gamma_N < 1$, Eq. (G44) tends in magnitude to infinity for a nondegenerate input. If $\alpha_N + \gamma_N = 1$, its magnitude is bounded below by $|d_\star\zeta|a^2$, so in either case it cannot tend to zero. One can also identify the limits on the equality line explicitly: for $\alpha_N = 1$, Eq. (G12) gives $J_N \to G_0$; for $\alpha_N > 1$, $\beta_N \to 0$ and

$$\sup_{s \leq N} \|E_N^s u_\star - u_\star\| + \sup_{s \leq N} \|(E_N^*)^s v_\star - v_\star\| \to 0,$$

$$J_N \to a\, v_\star \otimes u_\star.$$

For example, this follows by telescoping from $\|E_N^s - I\|_{\mathrm{op}} \le e^{2a\beta_N} - 1$. Thus the equality-line output is $-d_\star \zeta \|G_0\|_{\mathrm{HS}}^2$ at $\alpha_N = 1$, and $-d_\star \zeta a^2$ at $\alpha_N > 1$. This one-step routine satisfies the bounded-control assumptions for every exponent considered here. □

Proof of the linear specializations of Claims 1–3. Set $k = 1$ in Appendix G.2 for initialization stability, training stability, and the feature properties at $(1, 0)$. Lemma 11 supplies the stronger equivalence: throughout $\alpha_N \ge 1$, nontriviality holds exactly when $\alpha_N + \gamma_N \le 1$. Together with $\gamma_N \ge 0$ this selects $(1, 0)$. The general response proposition proves simultaneous state and preactivation learning, and prefix noncancellation gives all-routine faithfulness and $\kappa_N = 0$. □

G.4 A BLOCK-DEPTH-2 COUNTEREXAMPLE OUTSIDE TRAINING STABILITY

Setup. We construct an initialization-stable but training-unstable $k = 2$ model whose output changes by a nonzero amount after two SGD updates. Take $a = \eta = 1$, inputs $\xi_0 = \xi_1 = 1$, losses $\ell_0(f) = \ell_1(f) = f$, and

$$\gamma_N < 0, \qquad \alpha_N = 1 - 3\gamma_N.$$

Then $\alpha_N \ge 1$, so initialization is stable, while $\gamma_N < 0$ makes training unstable and $\alpha_N + \gamma_N = 1 - 2\gamma_N > 1$. The simplest instance is $(\alpha_N, \gamma_N) = (4, -1)$. Write

$$s = s_N = N^{-\gamma_N} \to \infty, \qquad \beta = \beta_N = N^{1-\alpha_N} = s^{-3}.$$

After taking the width limit at fixed N, write the two matrix operators as C_t and D_t, corresponding to $\boldsymbol{W}_t^1$ and $\boldsymbol{W}_t^2$, and set

$$u = u_\star, \quad v = v_\star, \quad C = C_0, \quad D = D_0, \quad p = Cu, \quad q = D^* v, \quad R(y, z) = y \otimes z.$$

The circular initialization gives $\|u\| = \|v\| = \|p\| = \|q\| = 1$, with u, p in the input source and v, q in the orthogonal readout source. Hence $\langle p, q\rangle = \langle u, v\rangle = 0$. All remainder estimates below are in operator norm after taking the width first, uniformly as $N \to \infty$.

The exact shared-gradient sums in Eq. (G5) give

$$C_{t+1} - C_t = -\frac{s}{N} \sum_{r=0}^{N-1} R\left(D_t^* (A_t^*)^{N-1-r} v, A_t^r u\right),$$

$$D_{t+1} - D_t = -\frac{s}{N} \sum_{r=0}^{N-1} R\left((A_t^*)^{N-1-r} v, C_t A_t^r u\right),$$

$$A_t = I + \frac{\beta}{N} D_t C_t.$$

Updates. At initialization $\sup_{r \le N} \|A_0^r - I\|_{\mathrm{op}} = \mathcal{O}(\beta)$, also for adjoints. Replacing these powers by the identity therefore gives

$$C_1 = C - sR(q, u) + \mathcal{O}(s^{-2}),$$
$$D_1 = D - sR(v, p) + \mathcal{O}(s^{-2}).$$

The apparent quadratic term in their product is zero:

$$s^2 R(v, p) R(q, u) = s^2 \langle p, q\rangle R(v, u) = 0.$$

Hence $\|D_1 C_1\|_{\mathrm{op}} = \mathcal{O}(s)$ and $\sup_{r \le N} \|A_1^r - I\|_{\mathrm{op}} = \mathcal{O}(\beta s) = \mathcal{O}(s^{-2})$. Both power estimates follow from $\|(I + \beta B/N)^r - I\|_{\mathrm{op}} \le e^{\beta \|B\|_{\mathrm{op}}} - 1$. Since $\|C_1\|_{\mathrm{op}} + \|D_1\|_{\mathrm{op}} = \mathcal{O}(s)$, replacing the powers in the second shared-gradient sums by the identity incurs $\mathcal{O}(s^{-1})$ error in each averaged rank-one sum and thus $\mathcal{O}(1)$ in each update. Consequently

$$D_1^* v = q - sp + \mathcal{O}(s^{-2}), \qquad C_1 u = p - sq + \mathcal{O}(s^{-2}),$$

and

$$\begin{aligned} C_2 &= C_1 - sR(D_1^* v, u) + \mathcal{O}(1) \\ &= C - 2sR(q, u) + s^2 R(p, u) + \mathcal{O}(1), \\ D_2 &= D_1 - sR(v, C_1 u) + \mathcal{O}(1) \\ &= D - 2sR(v, p) + s^2 R(v, q) + \mathcal{O}(1). \end{aligned}$$

In D_2C_2, the fourth-order term vanishes because $R(v,q)R(p,u) = \langle q,p\rangle R(v,u) = 0$. The two cubic terms are

$$(-2sR(v,p))(s^2R(p,u)) = -2s^3\|p\|^2R(v,u),$$
$$(s^2R(v,q))(-2sR(q,u)) = -2s^3\|q\|^2R(v,u).$$

All other explicit terms have degree at most two in s; multiplying an $\mathcal{O}(1)$ matrix remainder by the other matrix's $\mathcal{O}(s^2)$ leading part also gives $\mathcal{O}(s^2)$. Thus

$$D_2C_2 = -4s^3R(v,u) + \mathcal{O}(s^2).$$

Output. Since $\beta s^3 = 1$ and $\beta\mathcal{O}(s^2) \to 0$,

$$\beta D_2C_2 \to -4R(v,u).$$

The second-step loop operator consequently has the form

$$A_2 = I + \frac{1}{N}\left(\beta D_2C_2\right), \qquad A_2^N \to \exp\left(-4R(v,u)\right).$$

Since $R(v,u)^2 = \langle u,v\rangle R(v,u) = 0$, the exponential is $I - 4R(v,u)$. Pairing with the two unit vacua gives

$$f_2(1) - f_0(1) \to -4.$$

Thus, as illustrated by the diamond in Figure 7 (b), the training-unstable region contains a fixed-k parameterization with $\alpha_N + \gamma_N > 1$ that is nevertheless nontrivial. The nontriviality statement in Claim 2′ must therefore remain restricted to the training-stable region.

H Proofs for the Infinite Loop-Depth Limit in the Linear Case

This appendix proves Claims 4–6 of Appendix B for linear networks trained with SGD. The main classification is Lemma 18: the scale $s_{N,L}$ of the direct update and the scale $\rho_{N,L}$ of the output response give distinct training-stability conditions, while nontriviality depends on $\rho_{N,L}$. At the four-parameter Loop-μP point in Claim 6, Lemma 15 constructs the ordered depth and loop limits, and Lemmas 19 and 20 establish feature learning and the two diversity laws. Appendix H.1 develops the dynamics and bounds; Appendix H.2 uses them to prove Claims 4–6 stated in Appendix B.

Model. There are L distinct residual blocks, each shared across N loops. The input, readout, and hidden matrices are mutually independent, with $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{V}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,n^{-2})$, and $(\boldsymbol{W}_0^l)_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,n^{-1})$ for $1 \le l \le L$. For a scalar input ζ,

$$\begin{aligned}
\boldsymbol{x}_t^{1,0}(\zeta) &= \boldsymbol{U}\zeta, \qquad \boldsymbol{x}_t^{r,0}(\zeta) = \boldsymbol{x}_t^{r-1,L}(\zeta) \quad (r \ge 2),\\
\boldsymbol{h}_t^{r,l}(\zeta) &= \boldsymbol{W}_t^l\boldsymbol{x}_t^{r,l-1}(\zeta),\\
\boldsymbol{x}_t^{r,l}(\zeta) &= \boldsymbol{x}_t^{r,l-1}(\zeta) + c\boldsymbol{h}_t^{r,l}(\zeta),\\
f_t(\zeta) &= \boldsymbol{V}^\mathsf{T}\boldsymbol{x}_t^{N,L}(\zeta).
\end{aligned}$$

Here $1 \le r \le N$ and $1 \le l \le L$. The input and readout matrices are frozen, while all layer matrices are trainable and updated simultaneously. For each layer, the complete gradient

$$\boldsymbol{G}_t^l = \nabla_{\boldsymbol{W}^l}\ell_t(f_t(\xi_t))$$

includes all N loops and is evaluated before the simultaneous update. With

$$\widetilde{\boldsymbol{G}}_t^l = nN^{\alpha_N-1}L^{\alpha_L}\boldsymbol{G}_t^l,$$

the SGD update is

$$\boldsymbol{W}_{t+1}^l = \boldsymbol{W}_t^l - \eta n^{-1}N^{-\gamma_N}L^{-\gamma_L}\widetilde{\boldsymbol{G}}_t^l = \boldsymbol{W}_t^l - \eta N^{\alpha_N-1-\gamma_N}L^{\alpha_L-\gamma_L}\boldsymbol{G}_t^l.$$

Scales and limits. The constants $a, \eta > 0$ and the finite training horizon T are fixed. The ordered limits are $n \to \infty$ at fixed (L, N), then stack depth $L \to \infty$ at fixed N, and finally loop count $N \to \infty$. The exponents α_L, α_N control depth accumulation and loop reuse; γ_L, γ_N control the corresponding updates. A power $N^{u_N} L^{u_L}$ is bounded exactly when $(u_L, u_N) \preceq_{\mathrm{lex}} (0, 0)$; we use $\succeq_{\mathrm{lex}}$ for the reverse order.

Definition 9. Write

$$\begin{aligned} c &= aN^{-\alpha_N} L^{-\alpha_L}, \\ \beta_{N,L} &= N^{1-\alpha_N} L^{1/2-\alpha_L}, \\ s_{N,L} &= N^{-\gamma_N} L^{-\gamma_L}, \\ \rho_{N,L} &= N^{1-\alpha_N-\gamma_N} L^{1-\alpha_L-\gamma_L}, \\ b_{N,L} &= c\sqrt{L} = \frac{a}{N}\beta_{N,L}. \end{aligned} \tag{H1}$$

Thus $cs_{N,L} = a\rho_{N,L}/(NL)$. Here c is the branch multiplier, $\beta_{N,L}$ measures the change from the input at initialization, $s_{N,L}$ the direct effect of an update, $\rho_{N,L}$ its effect on the output, and $b_{N,L}$ the size of products within one cycle.

Training routine. The routines have bounded scalar inputs and realized loss derivatives, with simultaneous SGD updates over the fixed horizon. In singleton notation, $\Xi_t = (\xi_t)$, $\chi_t = \ell_t'(f_t(\xi_t))$, and $d_t = \eta\chi_t\xi_t$. Estimates are uniform on compact control sets and transfer to tight realized controls by truncation and subsequences. First-step lower bounds use $\xi_0 \neq 0$, $\chi_0 \to \chi_\star \neq 0$, and $d_\star = \eta\chi_\star\xi_0$, independently of L, N.

For a fixed finite batch $\Xi_t = (\xi_{t,b})_{b=1}^B$, linearity gives $f_t(\xi) = \xi f_t(1)$ and hence

$$\nabla_\theta \ell_t = \left(\sum_{b=1}^B \xi_{t,b}\, \partial_b \ell_t\right) \nabla_\theta f_t(1), \qquad d_t = \eta \sum_{b=1}^B \xi_{t,b}\, \partial_b \ell_t.$$

Here θ denotes one layer matrix or the full parameter list, and each $\partial_b \ell_t$ is evaluated at the current batch outputs. Fixed evaluation batches use $\|\cdot\|_{\mathrm{rms}}$; upper bounds assume bounded inputs, while lower bounds use a declared nondegenerate batch and the anti-concentration condition in Eq. (F1).

Colored representation. Put $\boldsymbol{u} = \boldsymbol{U}$ and $\boldsymbol{v} = n\boldsymbol{V}$. A rank-one operator is $(z \otimes w)y = z\langle w, y\rangle$; its finite-width counterpart is $\boldsymbol{z}\boldsymbol{w}^{\mathsf{T}}/n$. The color space H is $\mathbb{R}^L$ before the layer limit and $L^2[0,1]$ afterward. Each source space $\mathcal{H}_u, \mathcal{H}_v$ is a copy of $\mathcal{F}(H \oplus H) = \bigoplus_{m \geq 0} (H \oplus H)^{\otimes m}$, with empty-word unit vectors $u_\star, v_\star$. Creation prepends a letter; its adjoint contracts the first letter and kills the vacuum. As in Definition 8, define

$$C(f) = \mathsf{L}(f, +) + \mathsf{L}(f, -)^*, \qquad f \in H. \tag{H2}$$

The operators act identically on both source copies and satisfy $\|C(f)\|_{\mathrm{op}} \leq 2\|f\|$. Let $\mathcal{H}_0 \subset \mathcal{H}_u \oplus \mathcal{H}_v$ be the constant-color subspace generated by the constant functions in the two colors. Here $\mathbb{1} = \mathbb{1}_{[0,1]}$ is the constant function in $L^2[0,1]$. The operator $C(\mathbb{1})$ and its adjoint preserve $\mathcal{H}_0$; on $\mathcal{H}_0$, $C(\mathbb{1})$ is a one-color circular element.

Feature increments. For a fixed input put $\boldsymbol{y}_t^0 = \boldsymbol{U}\zeta$, $\boldsymbol{y}_t^r = \boldsymbol{x}_t^{r,L}(\zeta)$, and $r_N(\rho) = 1 + \lfloor \rho(N-1) \rfloor$. For $\rho, \lambda \in [0, 1)$ and sufficiently small $\varepsilon > 0$, the observables are

$$\begin{aligned} D_{N,t}(\rho, \varepsilon) &= \left\| \boldsymbol{y}_t^{\lfloor (\rho+\varepsilon)N \rfloor} - \boldsymbol{y}_t^{\lfloor \rho N \rfloor} \right\|_{\mathrm{rms}}, \\ D_{L,t}(\rho, \lambda, \varepsilon) &= N \left\| \boldsymbol{x}_t^{r_N(\rho), \lfloor (\lambda+\varepsilon)L \rfloor} - \boldsymbol{x}_t^{r_N(\rho), \lfloor \lambda L \rfloor} \right\|_{\mathrm{rms}}. \end{aligned}$$

The second increment is multiplied by N because it compares blocks within one cycle rather than complete loop endpoints. For a finite input batch, stack the state vectors and use $\|\cdot\|_{\mathrm{rms}}$. Upper estimates use bounded batches; lower estimates use the declared nondegenerate batch, with input magnitudes bounded away from zero. Each exponent uses the supremum and empty-set convention stated in Appendix A.1, with the network limits preceding $\varepsilon \downarrow 0$. Coordinatewise upper scales use tightness, while lower scales use the anti-concentration condition in Eq. (F1).

H.1 SGD TRAINING DYNAMICS

We first take the width limit at fixed (L, N), then establish the product and training bounds needed for the depth limit, and finally construct the depthwise continuum cycle and its loopwise limit at the distinguished point.

Widthwise limit. For fixed L, use the limiting representatives

$$\begin{aligned} \boldsymbol{W}_0^l &\longmapsto C(e_l) =: C_l, \\ \boldsymbol{u} &\longmapsto u_\star, \\ \boldsymbol{v} &\longmapsto v_\star, \\ \boldsymbol{z}\boldsymbol{w}^\mathsf{T}/n &\longmapsto z \otimes w, \end{aligned} \tag{H3}$$

where $e_1, \dots, e_L$ are orthonormal and $\longmapsto$ denotes representative correspondence rather than a mode of convergence.

Lemma 12. *Fix L, N, T. For mixed words P, Q and $\boldsymbol{e}, \boldsymbol{e}' \in \{\boldsymbol{u}, \boldsymbol{v}\}$,*

$$\frac{1}{n}\langle P(\boldsymbol{W}_0)\boldsymbol{e}, Q(\boldsymbol{W}_0)\boldsymbol{e}'\rangle \xrightarrow{\mathbb{P}} \langle P(C)e_\star, Q(C)e'_\star\rangle. \tag{H4}$$

The same Gram and scalar convergence holds for every fixed loop-depth SGD computation, uniformly on compact control sets. At tight realized controls, fixed coordinate rows converge along convergent control subsequences to a centered Gaussian mixture whose conditional covariance is the limiting Gram matrix; distinct rows are conditionally independent. Uniform Hilbert-norm upper bounds transfer to coordinatewise weak $\mathcal{O}(1)$ bounds, and positive lower bounds additionally give weak $\Theta(1)$ scales.

Proof. The proof is parallel to that of Lemma 4. The architectures are nevertheless different: a fixed-k residual branch composes k matrices inside one residual block, whereas a loop-depth cycle applies the L residual maps $I + c\boldsymbol{W}_0^l$ in order. We therefore give the L-colored width argument directly.

Consider a normalized trace of a mixed word of length $2m$ in the initial layer matrices and their transposes. Isserlis's formula pairs matrix entries, and independence of the layer matrices forces each pair to have the same layer color. The pairing identifies the ordered endpoints of the paired entries. Its connected quotient graph has at most m undirected edges and hence at most $m + 1$ free vertices, so its contribution has order n^{V-m-1}. Equality requires a tree with m distinct edges. The trace walk then visits every edge twice, once in each direction, so paired occurrences have opposite transpose signs as well as matching colors. Moreover, the walk exhausts each descendant component before returning across its parent edge, which makes the pairing noncrossing. Conversely, recursively removing an adjacent opposite-sign pair removes one leaf and one edge, proving that every noncrossing, opposite-sign, matching-color pairing has $m + 1$ free vertices and contributes one. All other pairings lose at least one power of n, and odd words have zero expectation.

For the variance, expand the product of two normalized traces. Pairings disconnected between the traces give the product of expectations. Every pairing connecting the two traces has a connected quotient graph with at most $m + 1$ free vertices, while the two-trace normalization is n^{-m-2}. Its contribution is therefore $\mathcal{O}(n^{-1})$. Since there are finitely many pairings at fixed word degree, normalized traces of any fixed finite family of mixed words converge jointly in L^2.

The vacuum moments of the colored circular operators obey the same rule. Expand each $C_l = \mathsf{L}(e_l, +) + \mathsf{L}(e_l, -)^*$ and its adjoint into creation and annihilation operators. A nonzero vacuum expectation matches creations with annihilations in nested or disjoint pairs; the creation–annihilation relations require matching colors, while the operator and adjoint occurrences have opposite signs. Thus the limiting trace moments are precisely the corresponding vacuum moments.

It remains to insert the independent input and readout vectors. Conditional on $\boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^L$, for every real matrix $\boldsymbol{M}$,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{u}/n \mid \boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^L] &= \operatorname{Tr}(\boldsymbol{M})/n, \\
\operatorname{Var}(\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{u}/n \mid \boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^L) &\le 2\|\boldsymbol{M}\|_{\mathrm{op}}^2/n, \\
\mathbb{E}[\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{v}/n \mid \boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^L] &= 0, \\
\mathbb{E}[|\boldsymbol{u}^\mathsf{T}\boldsymbol{M}\boldsymbol{v}/n|^2 \mid \boldsymbol{W}_0^1, \dots, \boldsymbol{W}_0^L] &= \|\boldsymbol{M}\|_{\mathrm{F}}^2/n^2 \le \|\boldsymbol{M}\|_{\mathrm{op}}^2/n.
\end{aligned}$$

A $1/4$-net of the unit sphere and a Gaussian union bound give, for fixed L,

$$\mathbb{P}\left(\max_{1\le l\le L} \left\|\boldsymbol{W}_0^l\right\|_{\mathrm{op}} > R\right) \le 2L\exp(2n\log 9 - nR^2/8).$$

Hence the operator norms of the finitely many initial matrices, as well as $\|\boldsymbol{u}\|_{\mathrm{rms}}$ and $\|\boldsymbol{v}\|_{\mathrm{rms}}$, have uniformly bounded fixed moments. Apply the displayed quadratic-form estimates with $\boldsymbol{M} = P(\boldsymbol{W}_0)^\mathsf{T} Q(\boldsymbol{W}_0)$. The fluctuations vanish in L^2, and the trace limits prove Eq. (H4), including orthogonality of the two source spaces.

Fix now L, N, T and a scalar control transcript. Every forward or backward vector is obtained by a finite sequence of matrix and transpose multiplications and linear combinations. Each complete layer gradient is a finite sum of normalized rank-one matrices, and all layer matrices are updated simultaneously. Induction over the fixed training horizon therefore expresses every computed vector as a finite linear combination of mixed-word vectors whose coefficients are polynomials in the controls and in finitely many basic Gram entries; normalized rank-one multiplication closes the induction through $(z \otimes w)y = z\langle w, y\rangle$. Every scalar contraction is a polynomial in the same quantities. The Gram convergence is consequently uniform on compact control sets. Compact truncation and subsequence extraction transfer it to tight realized controls without requiring continuity of the rule producing those controls.

Finally, simultaneous orthogonal conjugation of all L initial matrices and rotation of both source vectors preserves their joint law. The vector computations are equivariant, while the scalar controls are invariant. Apply an independent Haar rotation $\boldsymbol{Q}$ to the entire computation. Conditional on the unrotated computation, (ii) gives the Gaussian row limit determined by the normalized Gram matrix, uniformly on bounded Gram sets and including singular limits. Compact-uniform Gram convergence and convergent control subsequences therefore give the stated centered Gaussian mixture and conditional independence of distinct rows. The exact random-rotation moments in (iii) transfer uniform Hilbert-norm upper bounds to coordinate tightness, while positive limiting variances and Eq. (F2) give Eq. (F1). These arguments take $n \to \infty$ at fixed L, N, T; they give neither finite-width operator-norm convergence nor a width-convergence rate uniform in L, and they do not interchange the ordered limits. □

Depthwise limit.

Lemma 13. *A homogeneous polynomial of degree k in the colored circular operators, with coefficient tensor A, has norm at most $(k+1)\|A\|_{\mathrm{HS}}$. Every within-cycle segment of $I + cC_l$ is therefore bounded by*

$$F(b_{N,L}) := \sum_{k\ge 0} \frac{(k+1)b_{N,L}^k}{\sqrt{k!}}, \qquad b_{N,L} = c\sqrt{L}. \tag{H5}$$

Moreover, $F(b) = 1 + \mathcal{O}(b)$ as $b \downarrow 0$, and every segment spanning at most N loops is bounded by $F(b_{N,L})^{N+2}$. This bound is uniform when $\alpha_L > 1/2$, or when $\alpha_L = 1/2$ and $\alpha_N \ge 1$; at fixed N it is $1 + o_L(1)$ when $\alpha_L > 1/2$.

Proof. An annihilation of a minus letter followed by a creation of a plus letter is zero. Thus the expansion of an analytic word of length k has only $k+1$ nonzero types: j creations followed by $k-j$ annihilations, for $0 \le j \le k$. For a fixed type, flatten the coefficient tensor into a matrix from the annihilated initial tensor block to the created tensor block (reverse the annihilated indices). On each possible remaining suffix, the operator is this matrix tensored with the identity. Its norm is bounded by the matrix's Hilbert–Schmidt norm $\|A\|_{\mathrm{HS}}$. Summing the types proves the estimate. It extends by density to coefficient tensors over any real Hilbert space. The degree-k term of a segment

with $m \le L$ layers has $\binom{m}{k}$ distinct orthonormal word coefficients, each c^k. Its coefficient norm is $c^k\sqrt{\binom{m}{k}} \le b_{N,L}^k/\sqrt{k!}$, proving Eq. (H5). Split a longer segment at cycle boundaries. When $\alpha_L > 1/2$, $b_{N,L} \to 0$ first for every fixed N, so the ordered bound is one. When $\alpha_L = 1/2, \alpha_N \ge 1$, $b_{N,L} = aN^{-\alpha_N}$ and $Nb_{N,L} \le a$. The bound follows from $\log F(b_{N,L}) = \mathcal{O}(b_{N,L})$ near zero and the finitely many remaining N. □

For the width-limit computation, let $z_{t,r,l}$ be the forward vector at input one immediately before block application (r, l), and let $w_{t,r,l}$ be the transpose product of all later residual maps applied to $v_\star$.

Lemma 14. *In the initialization-stable region, the layer operators and responses satisfy exactly*

$$\begin{aligned} W_t^l &= C_l + s_{N,L}K_t^l, \qquad K_0^l = 0, \\ J_t^l &= \frac{a}{N}\sum_{r=1}^{N} w_{t,r,l} \otimes z_{t,r,l}, \\ K_{t+1}^l &= K_t^l - d_t J_t^l, \end{aligned} \tag{H6}$$

and every residual map is

$$I + cC_l + \frac{a\rho_{N,L}}{NL}K_t^l. \tag{H7}$$

If $\rho_{N,L}$ is bounded, then K_t^l, J_t^l and all forward and backward paths are uniformly bounded over depth and loop positions, with K_t^l, J_t^l bounded in Hilbert–Schmidt norm, and state and output changes are $\mathcal{O}(\rho_{N,L})$. If $s_{N,L}$ is also bounded, then preactivations and their changes are bounded. If $\rho_{N,L} \to 0$, every fixed SGD routine is trivial.

Proof. The exact finite-width gradient, with $\boldsymbol{z}, \boldsymbol{w}$ denoting its finite-width versions, is

$$\boldsymbol{G}_t^l = \frac{c\chi_t\xi_t}{n}\sum_r \boldsymbol{w}_{t,r,l}\boldsymbol{z}_{t,r,l}^{\mathsf{T}}. \tag{H8}$$

The normalized update rule above gives $-d_t s_{N,L}(a/N)\sum_r \boldsymbol{w}\boldsymbol{z}^{\mathsf{T}}/n$, proving Eqs. (H6) and (H7). No block application has been removed from this expression. Here is a uniform perturbation estimate usable despite noncommutativity and weight sharing. Suppose all unperturbed subproducts are bounded by $M \ge 1$, and a perturbation R_j is inserted at each position of the product, with $\sum_j \|R_j\|_{\mathrm{op}} \le A$. Expand the perturbed product by the set of perturbed positions. Its terms with k insertions have total norm at most $M^{k+1}A^k/k!$, since the sum over increasing distinct indices is at most $A^k/k!$. The full product is at most Me^{MA}, and its difference from the unperturbed one is at most $M(e^{MA} - 1)$. The estimate also applies to every subproduct and to adjoints. In Eq. (H7), $A \le a\rho_{N,L}\max_l \|K_t^l\|_{\mathrm{op}}$. Lemma 13 supplies bounded M in the ordered limit. If R_t bounds $\max\|K_t^l\|_{\mathrm{HS}}$, the forward and backward norms are at most $M\exp(Ma\rho_{N,L}R_t)$, and hence

$$\begin{aligned} \max_l \|J_t^l\|_{\mathrm{HS}} &\le aM^2\exp(2Ma\rho_{N,L}R_t), \\ R_{t+1} &\le R_t + |d_t|aM^2\exp(2Ma\rho_{N,L}R_t). \end{aligned} \tag{H9}$$

Finite induction gives a finite bound on every compact control set when $\rho_{N,L}$ is bounded. These are bounds on a fixed number of steps only. Comparing the updated and previous paths by a product telescoping identity, the sum of the residual-map differences is at most $a\rho_{N,L}|d_t|\max\|J_t^l\|_{\mathrm{op}}$. Both sets of subproducts are already bounded, so state and output changes are $\mathcal{O}(\rho_{N,L})$. For preactivations use $W_t^l = C_l + s_{N,L}K_t^l$ and $\|C_l\|_{\mathrm{op}} \le 2$. This proves their boundedness and that of their changes when $s_{N,L}$ is bounded. Compact truncation and Lemma 12 handle all routines. Summing finitely many $\mathcal{O}(\rho_{N,L})$ output changes proves the last assertion. Note that this argument rescales the residual maps and therefore does not assume the unrescaled trained matrices bounded when $s_{N,L}$ is unbounded. □

Loopwise limit. At the distinguished point $(\alpha_N, \alpha_L; \gamma_N, \gamma_L) = (1, 1/2; 0, 1/2)$, take $H = L^2[0,1]$, set $e_{L,l} = \sqrt{L}\,\mathbb{1}_{((l-1)/L, l/L]}$, and write $C_l = C(e_{L,l})$. Interpolate the width-limit fields

in Eq. (H6) cellwise as $K_{t,L,N}(\lambda)$ and $J_{t,L,N}(\lambda)$. Denote the depth-limit segment propagator asserted below by $P_{t,N}(s_1, s_0)$, with $P_{t,N} = P_{t,N}(1, 0)$, and write $x_t^{r,l}$ for the width-limit Hilbert representative of $\boldsymbol{x}_t^{r,l}$. For a bounded operator k, define

$$\begin{aligned} B(k) &= C(\mathbb{1}) + k, \\ X_k(s) &= e^{asB(k)} u_\star, \\ Y_k(s) &= e^{a(1-s)B(k)^*} v_\star, \\ G(k) &= a \int_0^1 Y_k(s) \otimes X_k(s) \, \mathrm{d}s. \end{aligned}$$

Lemma 15. *At each fixed N, the fields $K_{t,L,N}, J_{t,L,N}$ converge uniformly in Hilbert–Schmidt norm to continuous fields $K_{t,N}, J_{t,N}$ as $L \to \infty$. The corresponding segment propagators converge uniformly in operator norm, and the forward and backward vectors converge uniformly in norm. There are bounded operators k_t such that, with $B_t = B(k_t)$, $X_t = X_{k_t}$, $Y_t = Y_{k_t}$, and $G_t = G(k_t)$,*

$$\sup_{\lambda \in [0,1]} \left(\|K_{t,N}(\lambda) - k_t\|_{\mathrm{HS}} + \|J_{t,N}(\lambda) - G_t\|_{\mathrm{HS}} \right) = \mathcal{O}(N^{-1}).$$

$$k_0 = 0, \qquad k_{t+1} = k_t - d_t G_t. \tag{H10}$$

Uniformly in $r \le N$ and $\lambda \in [0, 1]$,

$$\begin{aligned} \lim_{L \to \infty} x_t^{r, \lfloor \lambda L \rfloor}(1) &= P_{t,N}(\lambda, 0) P_{t,N}^{r-1} u_\star \\ &= X_t((r-1)/N) + \mathcal{O}(N^{-1}). \end{aligned} \tag{H11}$$

The corresponding later-map backward vector equals $Y_t((r-1)/N) + \mathcal{O}(N^{-1})$. Moreover

$$P_{t,N}(s_1, s_0) = I + \frac{a}{N} \left[C(\mathbb{1}_{(s_0, s_1)}) + \int_{s_0}^{s_1} K_{t,N}(s) \, \mathrm{d}s \right] + \mathcal{O}(N^{-2}), \tag{H12}$$

uniformly in the endpoints. All convergence and error estimates are uniform on compact control sets.

Proof. Layer limit at fixed N. A base map is $I + (a/N)C(\mathbb{1}_{((l-1)/L, l/L]})$, and the trained map adds $(a/(NL))K_{t,L,N}^l$. For the base product over an interval, its degree-j coefficient tensor is $(a/N)^j$ times the indicator of tuples whose cell indices are strictly ordered in the multiplication order. As $L \to \infty$, this tensor converges in $L^2([0,1]^j)$ to $(a/N)^j$ times the ordered-simplex indicator on that interval. The failure set is contained in the sets where two coordinates lie in the same cell or a coordinate lies in an endpoint cell; its measure is $\mathcal{O}_j(L^{-1})$, uniformly over endpoints. Its coefficient norm is at most $(a/N)^j/\sqrt{j!}$. The homogeneous norm estimate of Lemma 13 and the summable majorant $(j+1)(a/N)^j/\sqrt{j!}$ prove uniform operator-norm convergence of the series. Denote the limit by $S_{a/N}(s_1, s_0)$. It is norm-continuous in endpoints, satisfies the concatenation identity, and

$$\begin{aligned} S_{a/N}(s_1, s_0) &= I + \frac{a}{N} C(\mathbb{1}_{(s_0, s_1)}) + \mathcal{O}(N^{-2}), \\ \sup_{s_0 \le s_1} \left\| S_{a/N}(s_1, s_0) - I \right\|_{\mathrm{op}} &= \mathcal{O}(N^{-1}). \end{aligned} \tag{H13}$$

Endpoint continuity follows as well from convergence of simplex indicators and the same summable majorant. For a bounded continuous Hilbert–Schmidt field K, set $\tau_0 := s_0$ and define the drifted product by the absolutely norm-convergent insertion series

$$\begin{aligned} P_{a/N,K}(s_1, s_0) = S_{a/N}(s_1, s_0) \\ + \sum_{j \ge 1} \left(\frac{a}{N} \right)^j \int_{s_0 < \tau_1 < \cdots < \tau_j < s_1} S_{a/N}(s_1, \tau_j) K(\tau_j) \\ S_{a/N}(\tau_j, \tau_{j-1}) \cdots K(\tau_1) S_{a/N}(\tau_1, s_0) \, \mathrm{d}\tau_1 \cdots \mathrm{d}\tau_j. \end{aligned} \tag{H14}$$

If M bounds all base segments and R bounds K, the j-th summand is bounded by $M^{j+1}(aR/N)^j/j!$. The finite-layer insertion expansion converges to Eq. (H14): at each fixed number of insertions its sums are Riemann sums, with base segment endpoints shifted by at most one cell, and the factorial

bound controls the tail uniformly. This proves convergence whenever the discrete fields converge uniformly to K. It also proves joint norm-continuity in endpoints and in the field (sup norm), and uniformity on compact families of such fields. At time zero the field is zero, so this construction applies. Suppose it applies at time t. In Eq. (H6), each of the finitely many r-terms at fixed N is a product of convergent segment propagators and complete cycle maps applied to one of the two Fock vacua. As a function of the cell position it therefore converges uniformly to a continuous vector path. For bounded vectors, $\|z \otimes w - z' \otimes w'\|_{\mathrm{HS}} \le \|z - z'\|\|w\| + \|z'\|\|w - w'\|$. Consequently Eq. (H6) converges uniformly to a continuous Hilbert–Schmidt field at time $t + 1$. This proves the fixed-N assertion by finite induction. The same induction gives uniformity on compact controls, since all constructions are continuous and the bounds Eq. (H9) are uniform.

Loop limit. The bounds Eq. (H9), with $\rho_{N,L} = 1$ and $b_{N,L} = a/N$, are uniform in N. In Eq. (H14), the terms with at least two insertions are $\mathcal{O}(N^{-2})$. In the one-insertion term replace each base propagator by the identity, using Eq. (H13); its error is $\mathcal{O}(N^{-2})$. Together with Eq. (H13) this proves Eq. (H12), with a constant depending only on the uniform field bound. It requires no bound on derivatives of the field. We prove Eqs. (H10) and (H11) by induction on training time. Initially $K_{0,N} = k_0 = 0$. If $K_{t,N} = k_t + \mathcal{O}(N^{-1})$, Eq. (H12) gives

$$\begin{aligned} P_{t,N} &= I + (a/N)B_t + \mathcal{O}(N^{-2}), \\ P_{t,N}(s_1, s_0) &= I + \mathcal{O}(N^{-1}). \end{aligned} \tag{H15}$$

All operators are uniformly bounded. The elementary exponential Taylor estimate and telescoping products give, for every $r \le N$,

$$\left\| P_{t,N}^r - e^{arB_t/N} \right\|_{\mathrm{op}} = \mathcal{O}(N^{-1}). \tag{H16}$$

Explicitly the one-step difference from $e^{aB_t/N}$ is $\mathcal{O}(N^{-2})$, and there are at most N terms with uniformly bounded surrounding subproducts. Prefixes and suffixes in Eq. (H15), and their adjoints, then prove the forward and backward assertions in Eq. (H11), including the omission of the current block application in the backward product. Substitute these vector approximations in Eq. (H6). Their rank-one errors are $\mathcal{O}(N^{-1})$ after averaging. The remaining Riemann sum over $(r-1)/N$ approximates the integral in Eq. (H10) with error $\mathcal{O}(N^{-1})$, because the bounded-operator exponential paths are Lipschitz in s. Thus $J_{t,N}(\lambda) = G_t + \mathcal{O}(N^{-1})$ uniformly in λ, and Eq. (H6) gives the next assertion for $K_{t+1,N}$. This completes the induction and the proof of the ordered construction. For frozen controls all displayed errors are deterministic norm bounds, uniform on compact control sets. For controls having no unique limiting law, apply these estimates along subsequences of the tight control transcript, exactly as in Lemma 12. The depth limit is always taken before the loop limit; no interchange of L and N is used. □

Lemma 16. *For every finite real control transcript in Eq.* (H10)*,*

$$\begin{aligned} &B_t u_\star \neq 0, \\ &\quad 0 < m \le \|B_t X_t(s)\| \le M < \infty \quad (0 \le s \le 1), \end{aligned} \tag{H17}$$

where the two bounds are uniform on compact control sets. Also $\|X_t(s)\|$ is bounded above and away from zero on such sets.

Proof. The constant-color space $\mathcal{H}_0$ is preserved by $C(\mathbb{1})$ and its adjoint. Induction in Eq. (H10) keeps both paths in $\mathcal{H}_0$, while each k_t has range there and annihilates $\mathcal{H}_0^\perp$. Restricted to $\mathcal{H}_0$, the recursion is exactly Eq. (G11) with $k = 1$ and $C_1 = C(\mathbb{1})$. Lemma 7, with $j = 1$ and $j = 0$, gives the generator and state lower bounds, uniformly on compact controls. □

H.2 Classification of loop-depth parameterizations

In this subsection, Lemmas 17 and 18 classify the general exponents; Lemmas 19 and 20 establish the local features and anisotropic diversity at the distinguished point.

Lemma 17. *Initialization stability is equivalent to*

$$(1/2 - \alpha_L, 1 - \alpha_N) \preceq_{\mathrm{lex}} (0, 0). \tag{H18}$$

In this region the initialized output has width limit zero. Moreover, within this region, a nonzero limiting initialization loop-endpoint increment is possible only at $\alpha_L = 1/2, \alpha_N = 1$.

Proof. Under Eq. (H18), Eq. (H5) bounds all hidden states. The inequality $\|C_l\|_{\mathrm{op}} \le 2$ bounds all preactivations, and the scalar output is bounded by the forward and readout norms. Lemma 12 transfers these bounds to coordinates. The two source spaces are preserved by each initial map, so the output contraction is zero in the width limit. For necessity, every analytic word applied to $u_\star$ is a pure plus-creation word. The degree-one component of the final state is

$$Nc\sum_{l=1}^{L} C_l u_\star.$$

It is orthogonal to all other degrees and has squared norm N^2Lc^2. If $\alpha_L < 1/2$, this diverges at fixed N as $L \to \infty$. If $\alpha_L = 1/2, \alpha_N < 1$, it diverges in the final N-limit. The fixed-coordinate width law is centered Gaussian with at least this variance, hence escapes every fixed bounded interval. This proves Eq. (H18). If $\alpha_L > 1/2$, every subpath tends in operator norm to the identity at fixed N: apply Eq. (H5) to each one-cycle map, also bounding its difference from the identity by $F(b_{N,L}) - 1$, and telescope the finitely many loops. If $\alpha_L = 1/2, \alpha_N > 1$, the same telescoping bound is $\mathcal{O}(Nb_{N,L}) \to 0$. Thus all initialization endpoint increments vanish in these cases. The remaining case is analyzed explicitly below. □

Lemma 18. *Conditional on initialization stability, training stability holds exactly when both $s_{N,L}$ and $\rho_{N,L}$ are bounded in the ordered limit. Nontriviality holds exactly when $\rho_{N,L}$ does not tend to zero. The nondegenerate first-step signal specified above supplies every necessary lower bound.*

Proof. At initialization $z_{0,r,l}$ belongs to the first source space and $w_{0,r,l}$ to the second, so each J_0^l has the two-source triangularity used in Proposition 8: it maps the first space into the second and kills the second, while every initial map preserves both. Two update insertions in one product of first-step maps therefore give a zero term, even when the insertions come from different layers. The complete first-step response is exactly linear in the update, for every fixed L, N and signed $d_\star = \eta\chi_\star\xi_0$ after the width limit. Taking the output contraction and using Eq. (H8) yields

$$f_1(\zeta) - f_0(\zeta) = -d_\star\zeta\,\frac{\rho_{N,L}}{L}\sum_{l=1}^{L}\left\|J_0^l\right\|_{\mathrm{HS}}^2. \tag{H19}$$

For verification of the coefficient, the output differential with respect to a layer matrix is $c\sum_r w_r \otimes z_r = (cN/a)J_0^l$; pair it with $-d_\star s_{N,L}J_0^l$ and use $cs_{N,L}N/a = \rho_{N,L}/L$. Every initial forward or backward vector has vacuum coefficient one. Hence

$$\begin{aligned}\langle v_\star, J_0^l u_\star\rangle &= a,\\ a \le \left\|J_0^l u_\star\right\| &\le \left\|J_0^l\right\|_{\mathrm{HS}} \le C,\end{aligned} \tag{H20}$$

where the last bound holds in the initialization-stable ordered region by Eq. (H5). Consequently the factor multiplying $-d_\star\zeta\rho_{N,L}$ in Eq. (H19) lies between a^2 and a finite positive constant. If $\rho_{N,L}$ diverges, the response is unbounded; if it has monomial scale one, the response does not vanish. If $\rho_{N,L} \to 0$, Lemma 14 proves triviality for all routines. At the training input the coefficient is $-\eta\chi_\star\xi_0^2\rho_{N,L}$ times this positive average. This settles nontriviality and the necessity of the aggregate-response condition for stability. In the very first block application the input remains u at both training times. Its width-limit preactivation change from this first update is

$$\Delta h_0^{1,1}(\zeta) = -d_\star\zeta s_{N,L}J_0^1 u_\star. \tag{H21}$$

Its coordinate law is centered Gaussian with standard deviation at least $|d_\star\zeta|as_{N,L}$. If $s_{N,L}$ diverges, this escapes every bounded interval, whether divergence occurs at the L or the N stage. Thus the local-update condition is independently necessary. Sufficiency is Lemma 14. For these monomials, boundedness of $s_{N,L}$ and $\rho_{N,L}$ means precisely

$$\begin{aligned}(-\gamma_L, -\gamma_N) &\preceq_{\mathrm{lex}} (0,0),\\ (1-\alpha_L-\gamma_L, 1-\alpha_N-\gamma_N) &\preceq_{\mathrm{lex}} (0,0),\end{aligned} \tag{H22}$$

and nonvanishing of $\rho_{N,L}$ means that its exponent pair is $\succeq_{\mathrm{lex}} (0,0)$. Stability and nontriviality therefore force both response exponents to vanish. Substituting $\gamma_L = 1-\alpha_L, \gamma_N = 1-\alpha_N$ into Eqs. (H18) and (H22) gives exactly

$$\begin{cases}\alpha_L = 1/2, & \alpha_N \ge 1,\\ 1/2 < \alpha_L < 1, & \alpha_N \in \mathbb{R},\\ \alpha_L = 1, & \alpha_N \le 1.\end{cases} \tag{H23}$$

This exhausts all real exponent choices. □

Lemma 19. *At the distinguished point every fixed SGD routine is faithful on the nondegenerate input batch. A nondegenerate initial training signal gives first-step feature learning at every positive macroscopic loop position, for both hidden states and preactivations. At every finite training time,*

$$\boldsymbol{G}_t^l = \mathcal{O}(n^{-1}L^{-1/2}), \tag{H24}$$

uniformly over layer positions, with matching coordinatewise $\Theta(n^{-1}L^{-1/2})$ under the same initial signal at time zero, including diagonal coordinates.

Proof. First record a local consequence of the continuum construction. Uniformly in l, the unit vectors $e_{L,l}$ converge weakly to zero in $L^2[0,1]$. Indeed for any f, $|\langle e_{L,l}, f\rangle|^2 \le \int_{((l-1)/L,l/L]} |f(s)|^2 \,\mathrm{d}s$, whose supremum tends to zero by absolute continuity of the integral. For each compact set $\mathcal{C}$ of Fock vectors this implies

$$\sup_{z\in\mathcal{C},l} \|\mathsf{L}(e_{L,l},-)^* z\| \to 0. \tag{H25}$$

To justify the extension, first approximate a vector by a finite sum of finite-degree simple tensors; weak convergence proves the assertion on each such tensor. The annihilators have norm one, so approximation and a finite net of $\mathcal{C}$ give uniformity. For fixed N and compact controls, all limiting forward vectors entering a block form a compact set, because their paths are norm-continuous on a compact parameter set, and there are finitely many loops. The depth convergence in Lemma 15 is uniform. Thus Eq. (H25), $C_l = \mathsf{L}(e_{L,l},+) + \mathsf{L}(e_{L,l},-)^*$, and the isometry of creation by a unit vector give

$$\|C_l z_{t,r,l}\| - \|z_{t,r,l}\| \to 0 \quad (L\to\infty), \tag{H26}$$

uniformly in positions and controls. The local trained correction is $s_{N,L}K_t^l = L^{-1/2}K_t^l$, so its action vanishes in this limit. By Eq. (H11), the final ordered preactivation norm is therefore $|\zeta|\|X_t(\rho)\|$ at macroscopic loop position ρ. The upper and lower bounds following Eq. (H17), also uniform over all $\rho\in[0,1]$, and Lemma 12 prove faithfulness. This proof concerns the local layer operator; it does not identify a single preactivation with the aggregate generator B_tX_t. For the first update put $G = G_0$ and $d_\star = \eta\chi_\star\xi_0$. Restricting to $\mathcal{H}_0$ and applying Proposition 8 with $k=1$ gives, for every fixed $a,\eta>0$,

$$X_1(\rho) - X_0(\rho) = -d_\star a \int_0^\rho e^{a(\rho-s)C(\mathbb{1})} G e^{asC(\mathbb{1})} u_\star \,\mathrm{d}s. \tag{H27}$$

The state projection bound Eq. (G32) specializes to

$$|\langle v_\star, X_1(\rho) - X_0(\rho)\rangle| \ge |d_\star| a^2 \rho > 0 \qquad (\rho>0). \tag{H28}$$

The uniform norm bounds give a matching finite upper bound. Applying Eqs. (H25) and (H26) also to the difference of the two states entering the block shows that the local preactivation change has limiting norm $|\zeta|\|X_1(\rho)-X_0(\rho)\|$: its explicit local parameter correction again vanishes with $L^{-1/2}$. Lemma 12 now proves the required $\Theta(1)$ coordinate response at every fixed positive loop position for this initial training signal. For gradients, Eq. (H8) becomes

$$n\sqrt{L}\,(\boldsymbol{G}_t^l)_{ij} = \frac{a\chi_t\xi_t}{N}\sum_{r=1}^{N} (\boldsymbol{w}_{t,r,l})_i (\boldsymbol{z}_{t,r,l})_j. \tag{H29}$$

On compact control sets both vector norms are uniformly bounded by Eq. (H9). The random-rotation moment identity Eq. (G10) and Cauchy–Schwarz bound the expected absolute value of the averaged coordinate products, proving Eq. (H24) by Markov's inequality and compact truncation. For completeness the first-step lower bound is a coordinate statement, not only a Hilbert–Schmidt norm bound. At initialization the two source spaces are orthogonal, so the limiting coordinate forward and backward Gaussian families are independent, even for $i=j$. For a Hilbert–Schmidt operator $J:\mathcal{H}_u\to\mathcal{H}_v$, its associated bilinear Gaussian form is $Z(J) = \sum_{j,k}\langle f_j, Je_k\rangle H_j F_k$, where F_k, H_j are independent standard Gaussians and the two bases are orthonormal. The series converges in L^2, with $\mathbb{E}\left[Z(J)^2\right] = \|J\|_{\mathrm{HS}}^2$; in particular this association is an isometry for differences in J. For fixed L, N, the width law of Eq. (H29) is $\chi_\star\xi_0 Z(J_0^l)$. Lemma 15 gives uniform Hilbert–Schmidt convergence $J_0^l \to J_{0,N}(\lambda)$, and the subsequent loop limit gives $J_{0,N}(\lambda)\to G_0$, uniformly in λ.

Embed all these operators into the same two continuum source spaces to apply the bilinear isometry. Thus Eq. (H29) has the final limiting law $\chi_\star\xi_0 Z(G_0)$. It is nonzero almost surely: conditionally on F, it is a centered Gaussian with variance $(\chi_\star\xi_0)^2\|G_0F\|^2$; this is finite almost surely and positive almost surely, because at least one row of G_0 is nonzero. Such a row exists by $\langle v_\star, G_0u_\star\rangle = a$. A nonzero linear combination of independent Gaussians is zero with probability zero. This proves the weak Θ scale, uniformly over layer sequences. □

Lemma 20. *At the distinguished point, for every fixed SGD routine and each nonzero scalar input, the ordered limiting increments satisfy*

$$D_{N,t}(\rho,\varepsilon) = |\zeta|\|X_t(\rho+\varepsilon) - X_t(\rho)\| = \Theta(\varepsilon), \tag{H30}$$

and, more explicitly,

$$\begin{aligned} D_{L,t}(\rho,\lambda,\varepsilon)^2 &= a^2\zeta^2\left[\varepsilon^2\|B_tX_t(\rho)\|^2 + \varepsilon(1-\varepsilon)\|X_t(\rho)\|^2\right] \\ &= \Theta(\varepsilon). \end{aligned} \tag{H31}$$

Thus $\kappa_N = 0$, $\kappa_L = 1/2$; the same layer increment without the factor N tends to zero. Within the stable nontrivial phase Eq. (H23)*, this is the unique point with a nondegenerate initialization endpoint path and an attained rescaled layer-diversity exponent $\kappa_L = 1/2$.*

Proof. Endpoint convergence Eqs. (H11) and (H16) gives the first equality in Eq. (H30). Floors change the macroscopic loop time by $\mathcal{O}(N^{-1})$ and disappear before $\varepsilon \downarrow 0$. The norm Taylor expansion gives

$$X_t(\rho+\varepsilon) - X_t(\rho) = a\varepsilon B_tX_t(\rho) + \mathcal{O}(\varepsilon^2),$$

uniformly in positions and compact controls. By Eq. (H17) the leading coefficient is bounded above and away from zero, proving Eq. (H30). For a layer interval $I = (\lambda, \lambda+\varepsilon)$, apply Eq. (H12) to its forward input in loop $r_N(\rho)$, and then Eqs. (H10) and (H11). It follows, taking depth first, that the rescaled difference has the final Hilbert-space limit

$$a\zeta\,[C(\mathbb{1}_I) + \varepsilon k_t]X_t(\rho). \tag{H32}$$

Indeed, after the depth limit its expansion error is $N\,\mathcal{O}(N^{-2}) \to 0$, and its input differs from $X_t(\rho)$ by $\mathcal{O}(N^{-1})$. In this step ε is fixed. All layer floors are removed by the fixed-N uniform endpoint convergence before the factor N is allowed to tend to infinity. Put $g = \mathbb{1}_I - \varepsilon\,\mathbb{1}$. It is orthogonal to the constant function, and $\|g\|^2 = \varepsilon(1-\varepsilon)$. Since $X_t(\rho)$ lies in the constant-color Fock subspace $\mathcal{H}_0$, $\mathsf{L}(g,-)^*X_t(\rho) = 0$. Also $\mathsf{L}(g,+)X_t(\rho)$ is orthogonal to $\mathcal{H}_0$. Consequently Eq. (H32) is the orthogonal sum

$$a\zeta\,[\varepsilon B_tX_t(\rho) + \mathsf{L}(g,+)X_t(\rho)]\,. \tag{H33}$$

Creation satisfies $\|\mathsf{L}(g,+)z\| = \|g\|\|z\|$, proving Eq. (H31). In particular its square-root lower bound uses the positive norm of X_t and cannot be canceled by training. At fixed ε, the unrescaled increment is $\mathcal{O}(N^{-1})$ after the depth limit by Eq. (H12), so it vanishes in the final limit. The two increment estimates establish the attained exponents $\kappa_N = 0$ and $\kappa_L = 1/2$. A proposed $\kappa_N > 0$ gives $D_{N,0}/\varepsilon^{1-\kappa_N} = \mathcal{O}(\varepsilon^{\kappa_N}) \to 0$; a proposed $\kappa_L > 1/2$ gives $D_{L,0}/\varepsilon^{1-\kappa_L} = \mathcal{O}(\varepsilon^{\kappa_L-1/2}) \to 0$ as $\varepsilon \downarrow 0$. Thus the initialization upper bounds exclude every larger exponent in the universal definitions. For uniqueness, Lemma 17 shows that elsewhere in Eq. (H23) the initialization endpoint increments vanish. The estimates at the distinguished point give the nondegenerate endpoint path and the attained layer exponent, proving their simultaneous uniqueness.

Batch extension. For a finite scalar input batch, every vector path is its path at input one times the corresponding scalar. Thus all squared batch norms in these arguments are multiplied by $\sum_j \zeta_j^2$, bounded above and away from zero for the declared batch. Coordinate scales follow componentwise. Uniform estimates on compact controls give the same weak laws for all routines by truncation. This completes the batch assertions. □

Proof of Claim 4. Lemma 17 proves sufficiency by the analytic product majorant and necessity by the orthogonal first-degree term of squared norm N^2Lc^2. □

Proof of Claim 5. Lemma 14 supplies universal finite-step upper bounds. The exact output and first-layer responses in Lemma 18 prove that bounded $\rho_{N,L}, s_{N,L}$ are necessary for training stability and that nonvanishing $\rho_{N,L}$ is equivalent to nontriviality under initialization stability. The same lemma gives the two lexicographic conditions and their three-region intersection. □

Proof of Claim 6. The initialization and anisotropic-increment results Lemmas 17 and 20 select $(1, \frac{1}{2}; 0, \frac{1}{2})$ within the stable nontrivial phase. Lemma 19 gives faithfulness, simultaneous state/preactivation learning, and the gradient scales. Lemma 20 gives $D_{N,t} = \Theta(\varepsilon)$, $D_{L,t} = \Theta(\sqrt{\varepsilon})$, their attained exponents, and the vanishing unscaled layer increment, including the batch and ordered-limit assertions. Substitution gives branch coefficient $a/(N\sqrt{L})$ and normalized SGD coefficient $\eta/(n\sqrt{L})$, hence raw SGD rate η. □

I Heuristics for the Proofs in the General Case

The linear analyses in Appendices G and H prove the claims stated in Appendices A and B only for networks trained with SGD, and this appendix justifies those claims for nonlinear networks and Adam. It develops the general-case proof mechanism for the one-block model under the assumptions in Appendix I.1. We establish the finite-network properties in Appendix I.2, then justify initialization stability in Appendix I.3, training stability and nontriviality in Appendix I.4, and feature properties in Appendix I.5. The final subsection records how the same mechanism extends to fixed block depth, loop-depth scaling, and entrywise optimizers.

I.1 Setup and assumptions

Model. The input, readout, and hidden matrices are mutually independent, with $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{V}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-2})$, and $\boldsymbol{W}_{0,ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-1})$. Put $\boldsymbol{v} = n\boldsymbol{V}$. Let $\mathbf{1}_n$ be the all-ones vector, $\boldsymbol{P}_n = \boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$, and $\mathrm{MS}(\boldsymbol{z}) = \boldsymbol{P}_n\boldsymbol{z}$. For a scalar input ζ the model is

$$\begin{aligned}
\boldsymbol{x}_t^0(\zeta) &= \boldsymbol{U}\zeta, \\
\boldsymbol{h}_t^r(\zeta) &= \boldsymbol{W}_t\boldsymbol{x}_t^{r-1}(\zeta), \\
\boldsymbol{x}_t^r(\zeta) &= \boldsymbol{x}_t^{r-1}(\zeta) + c_N\boldsymbol{P}_n\phi(\boldsymbol{h}_t^r(\zeta)), \qquad 1 \le r \le N, \\
f_t(\zeta) &= n^{-1}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{x}_t^N(\zeta).
\end{aligned}$$

The activation acts coordinatewise. The input and readout matrices are frozen, and only the shared hidden matrix is trained. For the complete shared gradient $\boldsymbol{G}_t = \nabla_{\boldsymbol{W}}\ell_t(f_t(\xi_t))$, the optimizer update is

$$\begin{aligned}
\widetilde{\boldsymbol{G}}_t &= nN^{\alpha_N - 1}\boldsymbol{G}_t, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta n^{-1}N^{-\gamma_N}Q_t(\widetilde{\boldsymbol{G}}_0, \dots, \widetilde{\boldsymbol{G}}_t).
\end{aligned}$$

The entrywise rule $Q_t(z_0, \dots, z_t) = z_t$ is SGD; other choices use the same external scales but need their own direction bounds and response conditions.

Scales and limits. The constants $a, \eta > 0$ and the finite training horizon T are fixed. We take $n \to \infty$ at fixed N, followed by $N \to \infty$. Write

$$\begin{aligned}
c_N &= aN^{-\alpha_N}, \qquad \beta_N = N^{1-\alpha_N}, \\
s_N &= N^{-\gamma_N}, \qquad \rho_N = \beta_N s_N.
\end{aligned}$$

Thus $Nc_N = a\beta_N$, s_N scales the direct parameter action, and ρ_N scales its accumulated residual response.

Training routine. The routine has bounded inputs and realized loss derivatives. Its singleton control is $\chi_t = \ell_t'(f_t(\xi_t))$. Estimates are first uniform on compact control sets and then transferred to realized controls by truncation. Nonzero initial training signals are required only for response lower bounds. Inputs are fixed across an update; upper bounds use fixed bounded batches, while lower bounds use a declared batch bounded away from zero. Finite-width vector and batch magnitudes use

$\|\cdot\|_{\mathrm{rms}}$, and coordinate lower scales use Eq. (F1). Unlike the linear Fock representation, the limiting X, H, R, Z below are scalar random variables of a joint coordinate law.

For a training batch, let $\chi_{t,b} = \partial_b \ell_t(\boldsymbol{f}_t(\Xi_t))$. The exact gradient is

$$\boldsymbol{G}_t = \sum_{b=1}^{|\Xi_t|} \chi_{t,b} \nabla_{\boldsymbol{W}} f_t(\xi_{t,b}), \qquad \boldsymbol{K}_{t,N} = \frac{a}{N} \sum_{b=1}^{|\Xi_t|} \sum_{r=1}^{N} \frac{\boldsymbol{z}_{t,b}^r (\boldsymbol{x}_{t,b}^{r-1})^{\mathsf{T}}}{n},$$

where $\boldsymbol{x}_{t,b}^r = \boldsymbol{x}_t^r(\xi_{t,b})$, the per-input backward recursion has terminal value $\boldsymbol{r}_{t,b}^N = \boldsymbol{v}\chi_{t,b}$, and $\boldsymbol{z}_{t,b}^r = \mathrm{diag}(\phi'(\boldsymbol{h}_{t,b}^r))\boldsymbol{P}_n \boldsymbol{r}_{t,b}^r$. Thus $\boldsymbol{G}_t = \beta_N \boldsymbol{K}_{t,N}$; for SGD, $\Delta \boldsymbol{W}_t = -\eta s_N \boldsymbol{K}_{t,N}$. Displayed single-input identities below describe each summand; the joint-limit hypotheses apply to the finite collection of inputs and their summed update. Nontriviality and feature-learning lower bounds use a singleton routine. Unlike in the linear model, no input-homogeneity reduction is used.

The strict linear results in Appendices G and H were proved without mean subtraction. The next lemma shows that inserting MS is asymptotically negligible for the observables used in those proofs, so their phase classifications also apply to the centered linear models.

Lemma 21 (Mean subtraction preserves the linear phase diagrams)**.** *For the identity activation, fix the block depth or stack depth, loop count, and training horizon in Appendices G and H. As $n \to \infty$, inserting* MS *changes every computed vector by $o_{\mathbb{P}}(1)$ in normalized RMS norm, and every normalized Gram, scalar contraction, and fixed coordinate by $o_{\mathbb{P}}(1)$, uniformly over bounded scalar update transcripts. Consequently, the linear initialization, training, feature, and diversity phase diagrams are unchanged by mean subtraction.*

Proof. Write $\boldsymbol{e} = \mathbf{1}_n/\sqrt{n}$ and $\boldsymbol{P}_n = \boldsymbol{I} - \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}$, and couple the two computations using the same Gaussian initialization and scalar update transcript. Every forward or backward vector $\boldsymbol{z}$ in the uncentered computation is orthogonally equivariant. Conditional on its orthogonally invariant data, Haar rotation gives $\mathbb{E}\,|\boldsymbol{e}^{\mathsf{T}}\boldsymbol{z}|^2 = \mathbb{E}\|\boldsymbol{z}\|_{\mathrm{rms}}^2$. At fixed structural parameters and training horizon, these norms are bounded on bounded-initialization events, uniformly over bounded transcripts. Hence

$$\|(\boldsymbol{I} - \boldsymbol{P}_n)\boldsymbol{z}\|_{\mathrm{rms}} = \frac{|\boldsymbol{e}^{\mathsf{T}}\boldsymbol{z}|}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0.$$

At fixed block depth or stack depth, loop count, and training horizon, each computation contains only finitely many vector operations and normalized rank-one updates. Inserting $\boldsymbol{P}_n$ therefore changes every computed vector by $o_{\mathbb{P}}(1)$ in normalized norm. The corresponding rank-one error is bounded by

$$\|\boldsymbol{z} - \boldsymbol{z}'\|_{\mathrm{rms}}\|\boldsymbol{w}\|_{\mathrm{rms}} + \|\boldsymbol{z}'\|_{\mathrm{rms}}\|\boldsymbol{w} - \boldsymbol{w}'\|_{\mathrm{rms}} = o_{\mathbb{P}}(1),$$

so normalized Grams and scalar contractions have the same limits. Permutation equivariance transfers the normalized mean-square estimate to each fixed coordinate. The bounds are uniform over bounded scalar update transcripts because these finite computations are polynomial in their scalar entries. These are precisely the observables used in the linear phase-diagram arguments of Appendices G and H. □

Assumption 1. The nonlinear arguments below use the assumptions listed in this display, with joint convergence over the fixed finite training and evaluation batches.

(i) **Regularity.** $\phi(0) = 0$, $\phi \in C^2$, and ϕ', ϕ'' are bounded. Products and response remainders used below have uniform fourth-moment control on bounded effective trajectories.

(ii) **Initialization closure.** At fixed N, the coordinate law satisfies

$$X^r = X^0 + c_N \sum_{j=1}^{r} \left(\phi(H^j) - \mathbb{E}\,\phi(H^j)\right),$$
$$X^0 \sim \mathcal{N}(0, q_0), \quad q_0 = \zeta^2,$$

where $(H^1, \ldots, H^N)$ is jointly centered Gaussian, independent of X^0, with $\mathbb{E}[H^r H^s] = \mathbb{E}[X^{r-1}X^{s-1}]$. Its activation covariance obeys

$$\mathrm{Cov}(\phi(A), \phi(B)) \geq b_\phi^2 c$$

for centered jointly Gaussian (A, B) of covariance $c \geq 0$, with a fixed $b_\phi > 0$.

(iii) **Trained closure.** Forward, backward, and response variables have compatible joint width limits, including correlations induced by training and transpose multiplication. On bounded effective trajectories their empirical inner products, products, and Riemann sums converge with the moment and compact-control uniformity needed for the integral equations below. Coordinates are exchangeable; normalized coordinate responses satisfy Eq. (F1), uniformly over the declared positions and compact control sets.

(iv) **Jacobian response.** For a fixed nondegenerate scalar training input with a finite nonzero initial loss-derivative limit and $\alpha_N \geq 1$, let $\bar{\boldsymbol{J}}_N(\boldsymbol{W}) = \beta_N^{-1} \nabla_{\boldsymbol{W}} f_N(\boldsymbol{W})$. At initialization its Frobenius norm has a finite positive limit, and fixed coordinates of $\bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\boldsymbol{x}^0$ have nonzero limits satisfying Eq. (F1). At $(1, 0)$, the residual-state and full preactivation directional responses to the same first SGD update are jointly nondegenerate on the declared inputs at every fixed positive loop fraction. The small-learning-rate remainder control is uniform enough to choose one fixed $\eta > 0$ for all prescribed positive loop fractions.

(v) **Trained nondegeneracy.** At $(1, 0)$, every training routine has uniformly nondegenerate preactivation coordinate laws and

$$0 < m_{t,\zeta} \leq \|\phi(H_t(\lambda, \zeta)) - \mathbb{E}\,\phi(H_t(\lambda, \zeta))\|_{L^2} \leq M_{t,\zeta} < \infty$$

uniformly over $\lambda \in [0, 1]$ and compact control sets.

(vi) **Integrated secant alignment.** For the full nontriviality classification with $\alpha_N \geq 1$ and $\rho_N \geq 1$, there is a fixed scalar training input whose initial loss derivative satisfies $\chi_0 \to \chi_\star > 0$, and a fixed $\eta > 0$ for which

$$\int_0^1 \langle \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0 - \theta\eta\chi_0 s_N \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)), \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\rangle_{\mathrm{F}}\, \mathrm{d}\theta$$

retains a positive lower bound in the ordered limit. The positive $\chi_\star$ preserves the direction of this secant hypothesis; it can be absorbed into the fixed positive step size after the width limit. At $(1, 0)$ the local Jacobian-response condition suffices in place of this condition.

Under Assumption 1, (i) and (ii) govern initialization stability; (i), (iii), (iv), and (vi) govern training stability and nontriviality; and (iii)–(v) govern faithfulness, feature learning, and diversity at the critical point.

Activation checks for Assumption 1. The following lemma verifies the activation regularity in Assumption 1 (i) and the covariance lower bound in (ii) for SiLU and GELU.

Lemma 22 (SiLU/GELU activation regularity and covariance)**.** *Suppose $\phi(z) = b_\phi z + e(z)$, where $b_\phi > 0$ and e is even and nondecreasing in $|z|$, with the required Gaussian moments. For centered jointly Gaussian (A, B) with covariance $c \geq 0$,*

$$\mathrm{Cov}(\phi(A), \phi(B)) \geq b_\phi^2 c.$$

Both $\mathrm{SiLU}(z) = z/(1 + e^{-z})$ and exact $\mathrm{GELU}(z) = z\Phi(z)$ satisfy this with $b_\phi = \frac{1}{2}$, where Φ is the standard Gaussian distribution function. They have bounded first and second derivatives, and

$$\mathbb{E}[\phi'(\sqrt{q}Z)] = \frac{1}{2},$$
$$\mathbb{E}[\phi'(\sqrt{q}Z)^2] \geq \frac{1}{4},$$
$$\mathrm{Var}(\phi(\sqrt{q}Z)) \geq \frac{q}{4}.$$

Proof. Joint sign symmetry kills the cross covariances between linear and even terms, leaving $b_\phi^2 c + \mathrm{Cov}(e(A), e(B))$. Conditional on $A = x$, B is Gaussian with fixed variance and a mean whose absolute value increases with $|x|$. For a standard Gaussian Z and $m, t \geq 0$,

$$\frac{\partial}{\partial m}\,\mathbb{P}(|m + \sigma Z| > t) = \frac{1}{\sigma}\left[\varphi((t - m)/\sigma) - \varphi((t + m)/\sigma)\right] \geq 0,$$

where φ is its density. Thus $\mathbb{E}\left[e(B) \mid A = x\right]$ increases with $|x|$, and its covariance with $e(A)$ is nonnegative. Degenerate conditional variances follow by continuity. Finally,

$$\mathrm{SiLU}(z) = \frac{z}{2} + \frac{z}{2}\tanh(z/2),$$
$$\mathrm{GELU}(z) = \frac{z}{2} + z(\Phi(z) - 1/2).$$

For $z \geq 0$, the derivatives of the even parts are $\frac{1}{2}\tanh(z/2) + \frac{z}{4}\operatorname{sech}^2(z/2)$ and $\Phi(z) - \frac{1}{2} + z\varphi(z)$, both nonnegative. Their first two derivatives are bounded. Symmetry gives the derivative mean, Jensen's inequality gives its second-moment lower bound, and the covariance bound with $A = B$ gives the variance inequality. □

Adam. For Adam, with zero initial moments,

$$\begin{aligned} m_t &= b_1 m_{t-1} + (1-b_1) z_t, \\ v_t &= b_2 v_{t-1} + (1-b_2) z_t^2, \\ Q_t &= \frac{m_t/(1-b_1^{t+1})}{\sqrt{v_t/(1-b_2^{t+1})} + \epsilon_{\mathrm{opt}}}, \end{aligned}$$

where $0 \leq b_1, b_2 < 1$ and $\epsilon_{\mathrm{opt}} > 0$ is fixed. The moments are updated separately at each parameter coordinate, using the normalized gradient history from Eq. (2).

I.2 Basic properties

This subsection records the finite-network identities and bounds used in the justifications below.

We first differentiate the finite shared computation before making any closure assumption. This keeps the sum over loops, the mean-subtraction projection, and the loss derivative explicit.

Lemma 23 (Exact shared nonlinear SGD)**.** *Along the training input, put*

$$\begin{aligned} \boldsymbol{r}_t^r &= n \nabla_{\boldsymbol{x}_t^r} \ell_t, \\ \boldsymbol{D}_t^r &= \operatorname{diag}(\phi'(\boldsymbol{h}_t^r)), \\ \boldsymbol{z}_t^r &= \boldsymbol{D}_t^r \boldsymbol{P}_n \boldsymbol{r}_t^r. \end{aligned}$$

Then

$$\begin{aligned} \boldsymbol{r}_t^N &= \boldsymbol{v} \chi_t, \\ \boldsymbol{r}_t^{r-1} &= \boldsymbol{r}_t^r + c_N \boldsymbol{W}_t^{\mathsf{T}} \boldsymbol{D}_t^r \boldsymbol{P}_n \boldsymbol{r}_t^r. \end{aligned} \tag{I1}$$

With normalized outer products represented as matrices, define

$$\boldsymbol{K}_{t,N} = \frac{a}{N} \sum_{r=1}^{N} \frac{\boldsymbol{z}_t^r (\boldsymbol{x}_t^{r-1})^{\mathsf{T}}}{n}. \tag{I2}$$

The exact identities are $\boldsymbol{G}_t = \beta_N \boldsymbol{K}_{t,N}$, $\widetilde{\boldsymbol{G}}_t = n \boldsymbol{K}_{t,N}$, and

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta s_N \boldsymbol{K}_{t,N}.$$

Proof. The state Jacobian in a loop is $\boldsymbol{I} + c_N \boldsymbol{P}_n \boldsymbol{D}_t^r \boldsymbol{W}_t$, whose transpose gives Eq. (I1). Differentiating every occurrence of $\boldsymbol{W}_t$ gives $(c_N/n) \sum_r \boldsymbol{z}_t^r (\boldsymbol{x}_t^{r-1})^{\mathsf{T}}$. Since $N c_N = a \beta_N$, this equals $\beta_N \boldsymbol{K}_{t,N}$. Substitution into the optimizer update with $Q_t(z_0, \ldots, z_t) = z_t$ gives the update. The loss derivative is already included in $\boldsymbol{r}_t^r, \boldsymbol{z}_t^r, \boldsymbol{K}_{t,N}$. In particular, at $(1, 0)$ the raw SGD update is $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \boldsymbol{G}_t$. □

Lemma 24 (Effective variables and finite-step bounds)**.** *Under Assumption 1* (i)*, set $\boldsymbol{B}_{t,N} = \beta_N \boldsymbol{W}_t$ and $\psi_\beta(z) = \beta \phi(z/\beta)$. Exactly,*

$$\begin{aligned} \boldsymbol{x}_t^r &= \boldsymbol{x}_t^{r-1} + \frac{a}{N} \boldsymbol{P}_n \psi_{\beta_N}(\boldsymbol{B}_{t,N} \boldsymbol{x}_t^{r-1}), \\ \boldsymbol{h}_t^r &= \beta_N^{-1} \boldsymbol{B}_{t,N} \boldsymbol{x}_t^{r-1}, \\ \boldsymbol{r}_t^{r-1} &= \boldsymbol{r}_t^r + \frac{a}{N} \boldsymbol{B}_{t,N}^{\mathsf{T}} \operatorname{diag}\left(\phi'(\boldsymbol{B}_{t,N} \boldsymbol{x}_t^{r-1}/\beta_N)\right) \boldsymbol{P}_n \boldsymbol{r}_t^r, \\ \boldsymbol{B}_{t+1,N} &= \boldsymbol{B}_{t,N} - \eta \rho_N \boldsymbol{K}_{t,N}. \end{aligned} \tag{I3}$$

For $\alpha_N \geq 1$, bounded ρ_N, bounded initialization norms, and a finite bounded control transcript, all $\boldsymbol{B}_{t,N}$, normalized state/backward norms, and $\|\boldsymbol{K}_{t,N}\|_{\mathrm{op}}$ are bounded uniformly in N. Moreover

$$\sup_{r \leq N} \|\Delta \boldsymbol{x}_t^r\|_{\mathrm{rms}} = \mathcal{O}(\rho_N), \qquad |\Delta f_t| = \mathcal{O}(\rho_N). \tag{I4}$$

If also $\gamma_N \geq 0$, then $\|\boldsymbol{W}_t\|_{\mathrm{op}}$ is bounded and $\sup_r \|\Delta \boldsymbol{h}_t^r\|_{\mathrm{rms}} = \mathcal{O}(s_N)$. In the first loop, without any small-update approximation,

$$\Delta \boldsymbol{h}_t^1(\zeta) = -\eta s_N \boldsymbol{K}_{t,N} \boldsymbol{U} \zeta. \tag{I5}$$

Proof. The rescaling is algebraic; in particular $\beta\phi(\boldsymbol{W}\boldsymbol{x}) = \psi_\beta(\boldsymbol{B}\boldsymbol{x})$. Uniformly in $\beta > 0$, $\psi_\beta(0) = 0$ and $|\psi'_\beta| \leq L_\phi$. On a bounded effective-operator ball this gives

$$\|\boldsymbol{x}_t^r\|_{\mathrm{rms}} \leq e^{aL_\phi \|\boldsymbol{B}_{t,N}\|_{\mathrm{op}}} \|\boldsymbol{U}\xi_t\|_{\mathrm{rms}},$$
$$\|\boldsymbol{r}_t^r\|_{\mathrm{rms}} \leq e^{aL_\phi \|\boldsymbol{B}_{t,N}\|_{\mathrm{op}}} \|\boldsymbol{v}\chi_t\|_{\mathrm{rms}},$$

and

$$\|\boldsymbol{K}_{t,N}\|_{\mathrm{op}} \leq \frac{aL_\phi}{N} \sum_{r=1}^{N} \|\boldsymbol{r}_t^r\|_{\mathrm{rms}} \|\boldsymbol{x}_t^{r-1}\|_{\mathrm{rms}}.$$

For a training batch, sum the kernel bound over its fixed number of inputs; bounded inputs and loss derivatives give the same size-independent upper bound. Finite induction bounds the effective recursion. Comparing two state trajectories and applying the discrete Gronwall inequality gives Eq. (I4); the readout satisfies $|\boldsymbol{V}^\mathsf{T}\boldsymbol{z}| \leq \|\boldsymbol{v}\|_{\mathrm{rms}} \|\boldsymbol{z}\|_{\mathrm{rms}}$. When s_N is bounded, induction also bounds $\boldsymbol{W}_t$ directly. Now

$$\Delta \boldsymbol{h}_t^r = \boldsymbol{W}_{t+1} \Delta \boldsymbol{x}_t^{r-1} + \Delta \boldsymbol{W}_t \boldsymbol{x}_t^{r-1}$$

has norm $\mathcal{O}(\rho_N + s_N) = \mathcal{O}(s_N)$ because $\beta_N \leq 1$. At $r = 1$ the input is unchanged, giving Eq. (I5). These are deterministic norm estimates on the specified events. □

Under Assumption 1 (iii), permutation equivariance and moment truncation transfer the vector bounds to coordinate tightness. For matrix entries, the product-moment bound in Eq. (I2) gives $(\boldsymbol{K}_{t,N})_{ij} = \mathcal{O}(n^{-1})$, hence $(\boldsymbol{G}_t)_{ij} = \mathcal{O}(n^{-1}\beta_N)$. Coordinate nonlinearities use permutation symmetry here.

I.3 Justification of initialization stability

Initialization classification. Under Assumption 1 (ii), a direct discrete argument applies for $\alpha_N < 1$. The coordinate-law closure is supplied by (ii); for SiLU and GELU, Lemma 22 verifies the covariance lower bound used below. Put $C_{r,s} = \mathbb{E}[X^r X^s]$ and denote the centered Gaussian activation covariance by $\mathcal{V}_\phi(q, q', c)$. Then

$$C_{r,s} = q_0 + c_N^2 \sum_{i=1}^{r} \sum_{j=1}^{s} \mathcal{V}_\phi(C_{i-1,i-1}, C_{j-1,j-1}, C_{i-1,j-1}). \tag{I6}$$

Induction using the kernel lower bound yields $C_{r,s} \geq q_0$ and

$$C_{N-1,N-1} \geq q_0 + b_\phi^2 q_0 c_N^2 (N-1)^2.$$

Since H^N is Gaussian with this variance after the fixed-N width limit, its coordinates escape bounded intervals when $\alpha_N < 1$. If $0 < \alpha_N < 1$, a local-time description uses $\tau = rc_N$:

$$\partial_\tau X_t = \mathcal{P}\phi(\mathcal{W}_t X_t),$$
$$-\partial_\tau R_t = \mathcal{W}_t^*[\phi'(\mathcal{W}_t X_t)\mathcal{P}R_t],$$
$$0 \leq \tau \leq a\beta_N.$$

The interval grows with N; gradients retain their exact sum form and need not remain bounded. For $\alpha_N \leq 0$ the same discrete covariance necessity applies, with no Euler step-size limit required. In the linear case $\phi(z) = b_\phi z$, the solution of Eq. (I6) is

$$C_{r,s} = q_0 \sum_{j=0}^{\min(r,s)} \binom{r}{j} \binom{s}{j} (b_\phi c_N)^{2j},$$

which recovers the one-block initialization calculation of Section 3.2.

Justification of Claim 1. Under Assumption 1 (i) and (ii), the initialization matrix norm and Lemma 24 bound the normalized states and preactivations; permutation symmetry gives coordinate tightness. The Gaussian preactivation law and Eq. (I6) exclude $\alpha_N < 1$. □

I.4 JUSTIFICATION OF TRAINING STABILITY AND NONTRIVIALITY

The branch case split remains internal to this subsection: the critical path is followed by the vanishing-residual path and the output-response calculation. The $\alpha_N < 1$ initialization instability was established in Appendix I.3 and is therefore excluded from the training analysis below. The existence of the joint limits and the passage from sums to integrals are supplied by Assumption 1 (i), (iii), (iv), and (vi).

Critical residual path.

Definition 10 (Shared limiting variables)**.** Under Assumption 1 (iii), work in the real L^2 space of the joint coordinate law, with $\langle Y, X\rangle = \mathbb{E}[YX]$, $\mathcal{P}X = X - \mathbb{E}\,X$, and $(Z \otimes X)Y = Z\,\mathbb{E}[XY]$. Write $\mathcal{W}_t, \mathcal{W}_t^*$ for the compatible actions representing forward and transpose computations, including their history-dependent correlations. Whenever the corresponding branch has a Riemann limit, define

$$\mathcal{K}_t = a\sum_{b=1}^{|\Xi_t|}\int_0^1 Z_{t,b}(s)\otimes X_t(s;\xi_{t,b})\,\mathrm{d}s,$$
$$Z_{t,b}(s) = \phi'(H_t(s;\xi_{t,b}))\mathcal{P}R_{t,b}(s), \qquad R_{t,b}(1) = v\chi_{t,b}.$$

Here U and v denote the scalar coordinate-limit variables of $\boldsymbol{U}$ and $\boldsymbol{v} = n\boldsymbol{V}$. For a singleton batch we suppress b and write Z_t, R_t, ξ_t, χ_t. The finite-N kernel is the corresponding sum over inputs of Eq. (I2).

For example, before the final branch limit the shared update acts on the entire trajectory as

$$\mathcal{W}_t X = \mathcal{W}_0 X - \eta a s_N \sum_{j<t}\sum_{b=1}^{|\Xi_j|}\int_0^1 Z_{j,b}(s)\langle X_j(s;\xi_{j,b}), X\rangle\,\mathrm{d}s,$$

and its adjoint exchanges the two vectors in each rank-one map. Consequently the covariance kernel $C_{t,j}(\lambda,\mu) = \mathbb{E}[X_t(\lambda)X_j(\mu)]$ depends on two loop positions. If $\mathcal{F}_t(\lambda) = \mathcal{P}\psi_{\beta_N}(\mathcal{B}_t X_t(\lambda))$, then

$$\partial_\lambda C_{t,j}(\lambda,\mu) = a\,\mathbb{E}[\mathcal{F}_t(\lambda)X_j(\mu)],$$
$$\partial_\lambda\partial_\mu C_{t,j}(\lambda,\mu) = a^2\,\mathbb{E}[\mathcal{F}_t(\lambda)\mathcal{F}_j(\mu)].$$

The expectations use the full trained joint law, including the shared initial operator.

Lemma 25 (Critical branch dynamics)**.** *Under Assumption 1* (i) *and* (iii)*, including joint convergence of the forward/backward computations and their product Riemann sums with the required moment bounds, at $\alpha_N = 1$ and $\gamma_N = 0$ the limiting SGD dynamics are*

$$\begin{aligned}
\partial_\lambda X_t &= a\mathcal{P}\phi(H_t),\\
H_t &= \mathcal{W}_t X_t,\\
X_t(0;\zeta) &= U\zeta,\\
-\partial_\lambda R_t &= a\mathcal{W}_t^*[\phi'(H_t)\mathcal{P}R_t], \qquad R_t(1) = v\chi_t,\\
\mathcal{W}_{t+1} &= \mathcal{W}_t - \eta\mathcal{K}_t,\\
f_t(\zeta) &= \langle v, X_t(1;\zeta)\rangle.
\end{aligned} \tag{I7}$$

For a training batch, the backward equation applies separately to each input and $\mathcal{K}_t$ is the summed kernel in Definition 10. At $\alpha_N = 1, \gamma_N > 0$, state, preactivation, and output responses are $\mathcal{O}(s_N)$ and vanish over a fixed training horizon.

Proof. At $\alpha_N = 1$, $\beta_N = 1$. Sum the exact forward and backward recursions in Lemma 23 over loops. By (iii), their joint limits satisfy

$$X_t(\lambda;\zeta) = U\zeta + a\int_0^\lambda \mathcal{P}\phi(H_t(s;\zeta))\,\mathrm{d}s,$$
$$R_t(\lambda) = v\chi_t + a\int_\lambda^1 \mathcal{W}_t^*[\phi'(H_t(s))\mathcal{P}R_t(s)]\,\mathrm{d}s.$$

The convergence in (iii) sends $\boldsymbol{K}_{t,N}$ to $\mathcal{K}_t$ as defined in Definition 10. For $\gamma_N = 0$, the exact update therefore gives $\mathcal{W}_{t+1} = \mathcal{W}_t - \eta\mathcal{K}_t$; the readout contraction converges jointly as well. Differentiating the integral equations gives Eq. (I7), with the stated initial and terminal values. The regularity and moment bounds used here are those in (i) and (iii).

For $\gamma_N > 0$, Lemma 24 gives state, preactivation, and output changes of order s_N at each step, since $\rho_N = s_N$. Summing over a fixed finite number of steps proves that training updates vanish. The passage to the limiting equations uses (i) and (iii). □

Vanishing residual path. Under Assumption 1 (i) and (iii), the bounded-weight branch for a singleton training input and $\alpha_N > 1$, $\beta_N \to 0$, satisfies

$$\begin{aligned} X_t(\lambda;\zeta) &= U\zeta + \mathcal{O}_{L^2}(\beta_N), \\ H_t(\lambda;\zeta) &= \mathcal{W}_t U\zeta + \mathcal{O}_{L^2}(\beta_N), \\ R_t(\lambda) &= v\chi_t + o_{L^2}(1). \end{aligned}$$

Thus

$$\mathcal{K}_t = a\,(\phi'(\mathcal{W}_t U\xi_t)\mathcal{P}(v\chi_t)) \otimes (U\xi_t) + o(1).$$

The matrix can still change at order one when $\gamma_N = 0$, but the hidden-state path freezes and the output response vanishes. If $\gamma_N < 0$ and $\rho_N \le 1$, use the effective variables instead:

$$\begin{aligned} \partial_\lambda X_t &= a\mathcal{P}\psi_{\beta_N}(\mathcal{B}_t X_t), \\ -\partial_\lambda R_t &= a\mathcal{B}_t^*[\phi'(\mathcal{B}_t X_t/\beta_N)\mathcal{P}R_t], \\ \mathcal{B}_{t+1} &= \mathcal{B}_t - \eta\rho_N\mathcal{K}_t. \end{aligned}$$

For $1 - \alpha_N < \gamma_N < 0$, Eq. (I4) still gives output triviality, while Eq. (I5) and (iv) give hidden-update instability. At $\gamma_N = 1 - \alpha_N$, the effective update is finite and is described through the recession function $\phi_\infty(z) = \lim_{\beta\downarrow 0} \beta\phi(z/\beta)$ when it exists. Gate convergence in the backward equation is supplied by (iii), including any mass at a zero effective field. The first update uses the original gate $\phi'(\mathcal{W}_0 U\xi_0)$ even though $\mathcal{B}_0 \to 0$.

Output response. Bounds on dynamics do not by themselves show that training changes the output or an internal feature. We therefore express these changes as parameter responses: a local derivative controls sufficiently small learning rates, whereas a finite update requires an integrated secant.

Lemma 26 (Local variation and integrated response)**.** *At finite width, put $\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x}$. The flow $\partial_\lambda \boldsymbol{x} = a\boldsymbol{P}_n\phi(\boldsymbol{W}\boldsymbol{x})$ has parameter variation $\boldsymbol{y}_{\boldsymbol{A}} = \partial_\theta \boldsymbol{x}_{\boldsymbol{W}+\theta\boldsymbol{A}}|_{\theta=0}$ given by*

$$\begin{aligned} \partial_\lambda \boldsymbol{y}_{\boldsymbol{A}} &= a\boldsymbol{P}_n\boldsymbol{D}(\lambda)(\boldsymbol{W}\boldsymbol{y}_{\boldsymbol{A}} + \boldsymbol{A}\boldsymbol{x}), \\ \boldsymbol{y}_{\boldsymbol{A}}(0) &= 0, \\ \boldsymbol{D}(\lambda) &= \operatorname{diag}(\phi'(\boldsymbol{W}\boldsymbol{x}(\lambda))). \end{aligned}$$

If $\boldsymbol{T}_{\boldsymbol{W}}(\lambda, s)$ is its state propagator, then

$$\begin{aligned} \boldsymbol{y}_{\boldsymbol{A}}(\lambda) &= a\int_0^\lambda \boldsymbol{T}_{\boldsymbol{W}}(\lambda, s)\boldsymbol{P}_n\boldsymbol{D}(s)\boldsymbol{A}\boldsymbol{x}(s)\,\mathrm{d}s, \\ D_{\boldsymbol{W}}\boldsymbol{h}(\lambda)[\boldsymbol{A}] &= \boldsymbol{A}\boldsymbol{x}(\lambda) + \boldsymbol{W}\boldsymbol{y}_{\boldsymbol{A}}(\lambda). \end{aligned} \tag{I8}$$

For the finite-N scalar-output network evaluated at its training input, write $\chi_0 = \ell_0'(f_N(\boldsymbol{W}_0))$, $\boldsymbol{J}_N(\boldsymbol{W}) = \nabla_{\boldsymbol{W}} f_N(\boldsymbol{W})$, and $\bar{\boldsymbol{J}}_N = \beta_N^{-1}\boldsymbol{J}_N$. The first SGD step follows $\boldsymbol{W}_\theta = \boldsymbol{W}_0 - \theta\eta\chi_0 s_N\bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)$, so its exact output response is

$$f_N(\boldsymbol{W}_1) - f_N(\boldsymbol{W}_0) = -\eta\chi_0\rho_N\int_0^1 \langle\bar{\boldsymbol{J}}_N(\boldsymbol{W}_\theta), \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\rangle_{\mathrm{F}}\,\mathrm{d}\theta. \tag{I9}$$

Proof. Differentiation of the flow and variation of constants give Eq. (I8). Differentiating $\boldsymbol{W}\boldsymbol{x}$ includes both the direct $\boldsymbol{A}\boldsymbol{x}$ term and the state-feedback term $\boldsymbol{W}\boldsymbol{y}_{\boldsymbol{A}}$. For Eq. (I9), integrate

$$\partial_\theta f_N(\boldsymbol{W}_\theta) = -\eta\chi_0 s_N\langle\boldsymbol{J}_N(\boldsymbol{W}_\theta), \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\rangle_{\mathrm{F}}$$

and use $\rho_N = \beta_N s_N$. In particular the derivative at $\eta = 0$ is $-\chi_0\rho_N\|\bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\|_{\mathrm{F}}^2$. □

Under Assumption 1 (iii) and (iv), Eq. (I8) passes to the shared limiting variables. At $(1, 0)$ and $A = -\mathcal{K}_0$, the preactivation response is

$$\mathcal{H}_\lambda = -\mathcal{K}_0 X_0(\lambda) - a\mathcal{W}_0 \int_0^\lambda T_0(\lambda, s)\mathcal{P}\left(\phi'(H_0(s))\mathcal{K}_0 X_0(s)\right) \mathrm{d}s.$$

The response statement in (iv) makes this full expression nondegenerate and controls its remainder for a fixed small learning rate. At general exponents the secant formula retains the entire finite update. Its integrand has the exact two-position expansion

$$\langle \bar{\boldsymbol{J}}_N(\boldsymbol{W}_\theta), \bar{\boldsymbol{J}}_N(\boldsymbol{W}_0)\rangle_{\mathrm{F}} = \frac{a^2}{N^2}\sum_{r,q}\langle \boldsymbol{z}_\theta^r, \boldsymbol{z}_0^q\rangle_n \langle \boldsymbol{x}_\theta^{r-1}, \boldsymbol{x}_0^{q-1}\rangle_n,$$

$$\langle \boldsymbol{z}, \boldsymbol{w}\rangle_n = \frac{\boldsymbol{z}^\mathsf{T}\boldsymbol{w}}{n},$$

where the backward terminal value is $\boldsymbol{v}$. The finite-step lower bound uses (vi).

These calculations also give a small-residual first-step response. To display the coefficients, take unit initial loss signal $\chi_\star = 1$, without specifying the loss. Under Assumption 1 (ii)–(iv), with moment estimates uniform in a fixed sufficiently small a, let $u = U\xi$, $q_0 = \mathbb{E}\, u^2 > 0$, $G = \mathcal{W}_0 u$, and $D = \phi'(G)$. Then

$$X_0(\lambda) = u + \mathcal{O}_{L^4}(a\beta_N), \qquad R_0(\lambda) = v + \mathcal{O}_{L^4}(a\beta_N),$$

$$\mathcal{K}_0 = a(Dv)\otimes u + \mathcal{O}_{\mathrm{HS}}(a^2\beta_N),$$

$$\|\mathcal{K}_0\|_{\mathrm{HS}}^2 = a^2 q_0\, \mathbb{E}[D^2] + \mathcal{O}(a^3\beta_N).$$

Independence of v from the initialization forward variables gives the last identity. Also $\mathcal{K}_0 u = aq_0 Dv + \mathcal{O}_{L^2}(a^2\beta_N)$ and, at $\alpha_N = 1$, $\mathcal{H}_\lambda = -aq_0 Dv + \mathcal{O}_{L^2}(a^2)$ uniformly in λ. Thus the scalar derivative lower bound supplies nonzero leading terms for the same-input Jacobian and feature responses. Responses at different inputs use (iv).

For the finite-residual boundary $\alpha_N > 1, \gamma_N = 1 - \alpha_N$, SiLU and GELU have recession function $\phi_\infty(z) = z_+$, with

$$\sup_z |\beta\phi(z/\beta) - z_+| \le C\beta.$$

This follows because their differences from ReLU are bounded on the real line. Under Assumption 1 (ii), (iii), and (vi), for unit input $\xi_0 = 1$ and unit initial loss signal $\chi_\star = 1$, the limiting variables are independent $U, v, G \sim \mathcal{N}(0, 1)$ and

$$\mathcal{K}_0 = a(\phi'(G)v)\otimes U, \qquad \mathcal{B}_1 = -\eta a(\phi'(G)v)\otimes U.$$

Writing $D = \phi'(G)$, the limiting forward equation preserves $\mathbb{E}[UX_1(\lambda)] = 1$, hence

$$X_1(\lambda) = U + a\lambda\mathcal{P}(-\eta a Dv)_+,$$

$$f_1 - f_0 = a\,\mathbb{E}[v(-\eta a Dv)_+] = -\frac{\eta a^2}{4}.$$

Indeed $\mathbb{E}[v(cv)_+] = c/2$ for every real c, and $\mathbb{E}\, D = 1/2$. This is a finite first-step output response at $\rho_N = 1$, while the preactivation still has the factor β_N^{-1}. The general large-update region is governed by the integrated secant condition.

Justification of Claim 2. This justification uses Assumption 1 (i), (iii), (iv), and (vi). In the initialization-stable region, weights are bounded when $\gamma_N \ge 0$ and the response bounds in Lemma 24 apply. Conversely, for $\gamma_N < 0$, the direct response in Eq. (I5) is unbounded for the Jacobian witness, disproving universal training stability. If $\rho_N \to 0$, the effective-variable estimate gives output triviality for every training routine, including the negative-γ_N region with bounded effective operators. At $(1, 0)$ (iv) gives a nonzero response for one fixed small learning rate. In the remaining $\rho_N \ge 1$ region, (vi) and Eq. (I9) give nontriviality. This uses a finite-step identity rather than extrapolating its derivative at zero learning rate. □

I.5 JUSTIFICATION OF FEATURE PROPERTIES AND EXTENSIONS

The preceding subsections establish the finite-network bounds, limiting dynamics, and response formulas for the one-block general case. At its critical point, under Assumption 1 (iii)–(v), the joint state/preactivation response and uniform remainder in (iv) give both feature-learning bounds for one fixed small learning rate. The path convergence and nondegeneracy in (iii) and (v) give faithfulness and diversity, completing the justification of Claim 3.

Justification of Claim 3. Under Assumption 1 (iii)–(v), initialization requires $\alpha_N \geq 1$ and hidden-update stability requires $\gamma_N \geq 0$. Every point in this region except $(1, 0)$ has $\rho_N \to 0$, hence is trivial. The local output response gives nontriviality at $(1, 0)$. By (v), all-routine faithfulness holds. By (iii) and Lemma 25,

$$X_t(\lambda + \varepsilon) - X_t(\lambda) = a\varepsilon \mathcal{P}\phi(H_t(\lambda)) + o_{L^2}(\varepsilon).$$

The upper and positive lower bounds in (v) imply norm $\Theta(\varepsilon)$ on the declared inputs. Floors disappear after the ordered network limits and before $\varepsilon \downarrow 0$, so $\kappa_N = 0$, and any larger exponent violates the upper bound. The discrete path has total variation $\mathcal{O}(1)$ and a sum of squared increments $\mathcal{O}(N^{-1})$; together with the path convergence in (iii) this gives a finite-variation random loop trajectory. □

Fixed block depth. For the fixed-k model in Eqs. (A1) and (A2), each of the k matrices is shared across the same N loops and updated simultaneously. Fixed matrix products change constants but introduce no additional powers of N, so the relevant scales remain β_N, s_N, and ρ_N. The general-case extension requires matrix-indexed versions of Assumption 1 (ii)–(v), with joint closure and response at every internal position. In particular, a nonzero branch-output response does not by itself imply that every internal preactivation changes; the response statement must hold for each prefix of the residual branch. Under these hypotheses, the one-block initialization and response arguments give the general-case heuristic for Claims $2'$ and $3'$. The nontriviality classification remains restricted to the training-stable region, while the matrix-indexed path convergence gives faithfulness, internal-feature learning, and $\kappa_N = 0$ at $(\alpha_N, \gamma_N) = (1, 0)$.

Loop-depth scaling. For the repeated-stack model, replace the single residual block by L centered nonlinear residual blocks and reuse the stack over N loops. The extension requires block-indexed versions of Assumption 1 (ii)–(v), uniform in L, together with the ordered limits $n \to \infty$, then $L \to \infty$, then $N \to \infty$. The initialization accumulation, direct block update, and aggregate response then have scales $\beta_{N,L}$, $s_{N,L}$, and $\rho_{N,L}$: the first block in the first loop isolates the local response, while summing over all NL block applications gives the aggregate response. A depthwise closure that retains the independent block fluctuations and remains uniformly nondegenerate yields $\kappa_L = 1/2$; the subsequent loop continuum gives the smooth exponent $\kappa_N = 0$. This is the heuristic mechanism behind Claims 4–6. It is not an automatic consequence of Claim 3, because the additional depth limit and anisotropic increments require their own uniform closure.

Remark 2 (Adam and other entrywise optimizers)**.** A general Q_t requires joint closure and upper bounds for its normalized matrix direction as in (iii), together with the direction–Jacobian alignment in (iv) for the output or feature under study. Its coefficient is $\eta N^{-\gamma_N}$ in the one-block and fixed-k settings and $\eta N^{-\gamma_N} L^{-\gamma_L}$ in the loop-depth setting, giving $L^{-1/2}$ at Loop-μP. Scale bounds control the upper response, but a matching lower response requires this alignment.