WIDTH AND DEPTH LIMITS COMMUTE DURING TRAINING IN LINEAR RESIDUAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work showed that width and depth limits in residual networks with branches scaled with $1/\sqrt{depth}$ commute at *initialization* in the sense that no matter how those limits are taken (e.g. width first, then depth, or depth first, then width, or with a fixed width/depth ratio, etc), the network output converges to the same limit at initialization. In this paper, we extend this result to commutativity *during training* for linear residual networks: we show that under the same $1/\sqrt{depth}$ branch scaling, for any fixed number of gradient steps with gradient clipping, the network output converges in $\mathbb{L}_2$ as $\min\{width, depth\} \to \infty$, implying commutativity of the width and depth limits during training. We further show that such result does not hold without gradient clipping, and a weaker result can be obtained in that case under the weaker convergence in probability.

1 INTRODUCTION

Modern neural networks are increasingly scaled along multiple architectural dimensions. The rapid growth and success of large language and vision models has further motivated the study of neural network scaling limits as a way to understand how model behavior changes with scale (Kaplan et al., 2020; Zhai et al., 2022). Among these limits, width and depth play distinct roles: increasing width gives rise to asymptotic regimes such as neural tangent kernel and mean-field limits (Jacot et al., 2018; Mei et al., 2019; Yang & Hu, 2020), while increasing depth connects residual networks to many continuous-time dynamics (Haber & Ruthotto, 2018; Hanin & Nica, 2019; Hayou et al., 2021; Hayou, 2022; Yang et al., 2024; Peluchetti & Favaro, 2020; Marion et al., 2022; Chizat, 2026). Understanding how these two limits interact is therefore fundamental when both dimensions grow simultaneously.

A central question is whether the width n and depth L limits commute, that is, whether the same asymptotic behavior is obtained regardless of the order or relative rate at which the two dimensions scale. For residual networks with branches scaled by $L^{-1/2}$, Hayou & Yang (2023) showed that the width and depth limits commute at initialization, so the same limit is obtained along arbitrary joint width–depth scaling paths. This leaves open whether commutativity persists during training. At initialization, the residual weights are independent across layers, a property used in existing joint commutativity analysis (Hayou & Yang, 2023; Yang et al., 2024; Hayou, 2025). During training, gradient descent couples the weights across layers. As a result, the arguments developed for initialization output convergence cannot directly extend. It is therefore unclear whether the width and depth limits continue to commute under the same $L^{-1/2}$ branch scaling.

We first show that standard gradient descent does not preserve commutativity during training because the large depth diverges in $\mathbb{L}_2$ after a single step: for every fixed finite width n, the trained output converges in probability as $L \to \infty$, while the corresponding $\mathbb{L}_2$ depth limit fails under standard gradient descent. We then show that, under clipped gradient descent, the width and depth limits of the trained output commute in $\mathbb{L}_2$ for any fixed number of training steps, with convergence rates $n^{-1/2}$ in width and $L^{-1/2}$ in depth.

Our proof proceeds by taking the two limits in sequence. We first establish the depth limit at finite width by comparing the trained residual network with a continuous-depth model, with estimates uniform in width. We then take the infinite-width limit of this continuous-depth system and identify the same deterministic trained output.

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.

Related Work. Infinite-width limits have been studied through Gaussian-process, neural tangent kernel, and mean-field descriptions (Jacot et al., 2018; Mei et al., 2019; Neal, 2012; Lee et al., 2017). In parallel, infinite-depth limits of residual networks motivate continuous-time descriptions of network dynamics (Haber & Ruthotto, 2018; Hayou et al., 2021; Marion et al., 2022; Peluchetti & Favaro, 2020; Chizat, 2026), and more heuristic analyses of training dynamics (Yang et al., 2024; Bordelon et al., 2023). The interaction between these two limits was studied by Hayou (2023), who proved width–depth commutativity for the neural covariance of ResNets with residual branches scaled by $L^{-1/2}$. Hayou & Yang (2023) extended this analysis to more general residual scalings, while non-commutative limits have been identified for some MLP and ResNet parameterizations (Li et al., 2022; Noci et al., 2023). More recently, Chizat (2026) studied the large-depth training limit of ResNets scaled with L^{-1} branch multiplier at fixed network width. Chaintron et al. (2026) is probably the closest work to our paper. The authors showed that commutativity of the width and depth limits holds during training with L^{-1} branch scaling. Such scaling leads to ODE-style infinite-depth dynamics, which is substantially different from the SDE-style dynamics obtained with $L^{-1/2}$. Our work establishes commutativity of the width and depth limits during training under the $L^{-1/2}$ branch scaling consistent with Depth-μP (Yang et al., 2024). These two scaling choices lead to completely different dynamics and therefore require different proof machinery to establish commutative limits.

2 Setup and Notation

2.1 Model and parameterization

We consider a linear residual network with input dimension d, width n, and depth L. For a fixed nonzero input $x \in \mathbb{R}^d$, the hidden states and scalar output are given by

$$\begin{aligned} Y_0 &= W_0\, x, \\ Y_\ell &= Y_{\ell-1} + L^{-1/2}\, W_\ell\, Y_{\ell-1}, \qquad \ell = 1, \dots, L, \\ f_{n,L}(x) &= \frac{1}{n}\, \xi^\top Y_L, \end{aligned} \tag{1}$$

where $W_0 \in \mathbb{R}^{n\times d}$ is the input weight matrix, $W_\ell \in \mathbb{R}^{n\times n}$ for $\ell = 1, \dots, L$ are the residual weights, and $\xi \in \mathbb{R}^n$ is the output weight.[1] Since we will study both the width and depth limits, we parameterize some hyperparameters w.r.t to these scaling dimensions: we use the $L^{-1/2}$ branch multiplier to stabilize the depth limit, which was suggested by Depth-μP (Yang et al., 2024), and we also stabilize the width limit by randomly initializing the weights as follows: $(W_0)_{ij} \sim \mathcal{N}(0, d^{-1})$, $(W_\ell)_{ij} \sim \mathcal{N}(0, n^{-1})$, and $\xi_i \sim \mathcal{N}(0, 1)$, which also corresponds to μP (Yang & Hu, 2020).

To simplify the analysis, we keep W_0 and ξ fixed and train only the residual weights $W_1, \dots, W_L$. Our results can be extended to the case where W_0 and ξ are also trainable.[2] Throughout the paper, $\|z\|_2$ denotes the Euclidean norm of a vector z. For a random vector Z, we write $\|Z\|_{\mathbb{L}_2} := (\mathbb{E}\|Z\|_2^2)^{1/2}$; for a scalar random variable, the inner norm is understood as the absolute value. For $u, v \in \mathbb{R}^n$, set $\langle u, v\rangle_n := n^{-1} u^\top v$ and $\|u\|_n := \langle u, u\rangle_n^{1/2} = n^{-1/2}\|u\|_2$. We use the notation $q_0 := \|x\|_2^2/d$. Conditional on x, initial hidden state satisfies $Y_0 \sim \mathcal{N}(0, q_0 I_n)$.

2.2 Width–Depth Commutativity

We formalize width–depth commutativity following Hayou (2023). Unless stated otherwise, commutativity refers to convergence in $\mathbb{L}_2$ of the network output as $\min\{n, L\} \to \infty$.

Definition 1 (Width–depth commutativity)**.** Fix a training iteration k. We say that the width and depth limits commute if there exists a deterministic limit $f_\infty^{(k)}$ such that

$$\left\| f_{n,L}^{(k)} - f_\infty^{(k)} \right\|_{\mathbb{L}_2} \longrightarrow 0 \qquad \text{as } \min(n, L) \to \infty.$$

[1] While we consider a one-dimensional output, our results can be extended to linear residual networks with multi-dimensional outputs.

[2] We do not include such extension since it does not add any valuable insights to the proof machinery.

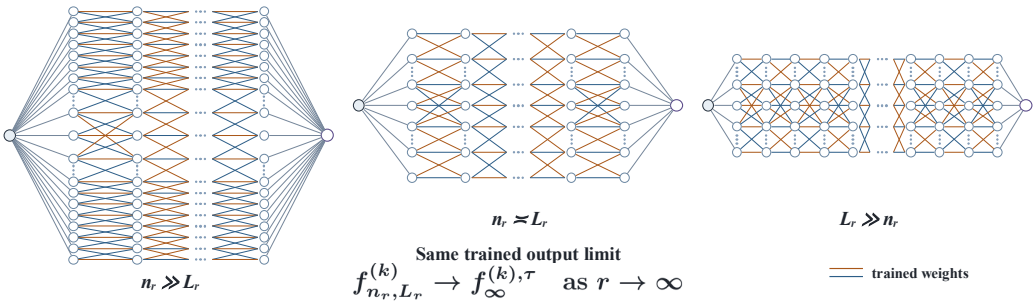


Figure 1: **Commutativity of the width and depth limits.** Illustration of three relative order regimes for (n_r, L_r) as $n_r, L_r \to \infty$. For the same fixed input and any fixed training iteration k, all three paths converge to the same trained output limit $f^{(k)}_{\infty}$.

We also consider a weaker notion of commutativity, requiring only convergence in probability. This will be useful because we will establish commutativity in probability but not in $\mathbb{L}_2$ in some cases.

Definition 2 (Commutativity in probability)**.** Fix a training iteration k. We say that the width and depth limits commute in probability if there exists a deterministic limit $f^{(k)}_{\infty}$ such that, for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|f^{(k)}_{n,L} - f^{(k)}_{\infty}\Big| > \varepsilon\Big) \longrightarrow 0 \qquad \text{as } \min(n, L) \to \infty.$$

Both notions require the limiting output to be independent of the relative growth of width and depth. Moreover, commutativity in $\mathbb{L}_2$ implies commutativity in probability.

2.3 Training dynamics

Without loss of generality, we train model (1) by minimizing a quadratic loss function evaluated on single input–target pair $(x, y_\star)$

$$\mathcal{L}(W_1, \dots, W_L) = \frac{1}{2}\big(f_{n,L}(x) - y_\star\big)^2.$$

This setup is used to alleviate the notation in the proofs, and it was used in Hayou (2025; 2023) for similar purposes. Our results can be readily extended to the case of a finite training dataset at the cost of redundant computations.

To minimize the loss, we update the weights using gradient descent (GD)

$$W^{(k+1)}_{\ell} = W^{(k)}_{\ell} - \eta \nabla_{W^{(k)}_{\ell}} \mathcal{L}, \qquad \ell = 1, \dots, L, \quad k = 0, \dots, T-1, \tag{2}$$

where $\eta > 0$ is the learning rate, fixed independently of n and L, and $W^{(0)}_{\ell}$ is given by the initialization above. These initialization rules combined with the depth multiplier above ($L^{-1/2}$) follow the scaling rules suggested by Yang & Hu (2020) and Yang et al. (2024) to guarantee stability and feature learning as model width n and depth L grow.

For a training step k, let $A^{(k)}_{\ell} := I + L^{-1/2} W^{(k)}_{\ell}$. The corresponding forward and backward states are defined by

$$\begin{aligned} &\textbf{Forward:} \quad Y^{(k)}_0 = Y_0, \quad Y^{(k)}_{\ell} = A^{(k)}_{\ell}\, Y^{(k)}_{\ell-1}, \qquad \ell = 1, \dots, L, \\ &\textbf{Backward:} \quad B^{(k)}_L = \xi, \quad B^{(k)}_{\ell-1} = \big(A^{(k)}_{\ell}\big)^\top B^{(k)}_{\ell}, \quad \ell = L, \dots, 1. \end{aligned} \tag{3}$$

Since we consider a single fixed training pair $(x, y_\star)$, we omit it from the notation and write the neural network output at training step k, $f^{(k)}_{n,L}$. We also write $\varepsilon^{(k)}_{n,L} := f^{(k)}_{n,L} - y_\star$. Straightforward calculations give the gradients of the network output w.r.t of the hidden weights $\nabla_{W^{(k)}_{\ell}} f^{(k)}_{n,L} = \frac{1}{n\sqrt{L}} B^{(k)}_{\ell} (Y^{(k)}_{\ell-1})^\top$.

3 Width and Depth Limits at $k = 1$

We first develop an intuitive understanding of the problem after a single gradient step. The width and depth limits commute in probability, but not in $\mathbb{L}_2$ (Definition 1, Definition 2).

Proposition 3.1 (Commutativity in probability after one GD step)**.** After one gradient step,

$$f_{n,L}^{(1)} \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{P}} f_{\infty}^{(1)} = \eta y_\star q_0 e,$$

The proof is given in Appendix D, which also establishes commutativity in probability after any fixed finite number of gradient steps. The next proposition shows that this commutativity does not extend to $\mathbb{L}_2$. In fact, after one gradient step, the depth limit already fails in $\mathbb{L}_2$ at every fixed finite width.

Proposition 3.2 (Depth-wise $\mathbb{L}_2$ divergence under gradient descent)**.** After the first gradient iteration, for every fixed finite $n \geq 1$, $\eta > 0$, and $y_\star \in \mathbb{R}$, $\lim_{L\to\infty} \mathbb{E}\left|f_{n,L}^{(1)}\right|^2 = +\infty$.

Together, Propositions 3.1 and 3.2 show that, after one gradient step, the width and depth limits commute in probability, while the depth-first $\mathbb{L}_2$ limit does not exist. Indeed, for each n one may choose $L_n \geq n$ sufficiently large so that $\mathbb{E}|f_{n,L_n}^{(1)}|^2 \geq n$. Then $\min\{n, L_n\} \to \infty$. Hence, an unrestricted joint $\mathbb{L}_2$ limit cannot hold under ordinary gradient descent. As shown in Appendix A, this failure arises from the unbounded first GD update as depth increases. This motivates the clipped gradient descent studied next. Specifically, we use clipped gradient descent with clipping radius $\tau > 0$. We denote the clipped training trajectory by the superscript τ. Define $J_{n,L}^{(k),\tau} := \frac{1}{L}\sum_{\ell=1}^{L} \|Y_{\ell-1}^{(k),\tau}\|_n^2 \|B_\ell^{(k),\tau}\|_n^2$. Let $\gamma_{n,L}^{(k),\tau} := \eta\varepsilon_{n,L}^{(k),\tau} \min\{1, \tau/\left(|\varepsilon_{n,L}^{(k),\tau}|\sqrt{J_{n,L}^{(k),\tau}}\right)\}$,[3] where $\varepsilon_{n,L}^{(k),\tau} := f_{n,L}^{(k),\tau} - y_\star$. The clipped updates are given by

$$W_\ell^{(k+1),\tau} = W_\ell^{(k),\tau} - \frac{\gamma_{n,L}^{(k),\tau}}{n\sqrt{L}} B_\ell^{(k),\tau}(Y_{\ell-1}^{(k),\tau})^\top.$$

By definition of $\gamma_{n,L}^{(k),\tau}$, $\left(\sum_{\ell=1}^{L} \|W_\ell^{(k+1),\tau} - W_\ell^{(k),\tau}\|_F^2\right)^{1/2} = |\gamma_{n,L}^{(k),\tau}|\sqrt{J_{n,L}^{(k),\tau}} \leq \eta\tau$. Thus, clipping enforces uniform control of each GD update, independently of n and L. The next section establishes width-depth commutativity in $\mathbb{L}_2$ under gradient clipping.

4 Width–Depth Commutativity during Training

We now state the main result. For any fixed number of training iterations, the output of a linear residual network trained with clipped gradient descent converges to the same deterministic limit along every joint path $(n, L) \to (\infty, \infty)$.

Theorem 4.1 (Commutative Width and Depth Limits during Training)**.** Fix $T > 0$ number of training steps, and $\tau > 0$. There exist deterministic limits $f_\infty^{(0),\tau}, \ldots, f_\infty^{(T),\tau}$ and a finite constant C_T, independent of n and L, such that

$$\max_{0\leq k\leq T} \left\| f_{n,L}^{(k),\tau} - f_\infty^{(k),\tau} \right\|_{\mathbb{L}_2} \leq C_T \left(n^{-1/2} + L^{-1/2}\right). \tag{4}$$

The bound in Theorem 4.1 holds along any sequence $n, L \to \infty$, thus implying the existence of the limit as $\min\{n, L\} \to \infty$ and therefore commutativity in $\mathbb{L}_2$.

Remark 4.2 (Sharpness in width)**.** The $n^{-1/2}$ term in Theorem 4.1 cannot be improved. At initialization ($k = 0$), independence of ξ and Y_L gives $\mathbb{E}|f_{n,L}^{(0)}|^2 = \frac{1}{n}\mathbb{E}\|Y_L^{(0)}\|_n^2 = \frac{q_0(1+L^{-1})^L}{n}$. Since

[3] If the denominator vanishes, the multiplier inside the minimum is set to one.

$(1 + L^{-1})^L \leq e$, we have $f_{n,L}^{(0)} \to 0$ in $\mathbb{L}_2$ as $n \to \infty$, uniformly in L. Hence, $f_{\infty}^{(0),\tau} = 0$. $\|f_{n,L}^{(0)} - f_{\infty}^{(0),\tau}\|_{\mathbb{L}_2} = \sqrt{\frac{q_0(1+L^{-1})^L}{n}} \geq \sqrt{\frac{q_0}{n}}$, establishing the sharpness of the $n^{-1/2}$ bound. The same cannot be said about the depth rate $L^{-1/2}$ as we will later provide empirical results showing that it is unlikely to be sharp.

4.1 Understanding the difficulty of proving commutativity during training

The initialization argument in Hayou & Yang (2023) does not extend directly after training because gradient descent couples the residual blocks across layers.

At initialization, $A_\ell^{(0)} = I + L^{-1/2}W_\ell$ for $\ell = 1, \dots, L$, where $\{W_\ell\}_{\ell=1}^L$ are independent Gaussian matrices. Let $t_\ell = \ell/L$. Since $L^{-1/2}W_\ell$ has the same Gaussian distribution as a matrix-Brownian increment over an interval of length $1/L$, one can construct a matrix Brownian motion M on the same probability space such that

$$M_{t_\ell} - M_{t_{\ell-1}} = L^{-1/2}W_\ell, \qquad \ell = 1, \dots, L.$$

Under this coupling,

$$Y_\ell^{(0)} = \big(I + M_{t_\ell} - M_{t_{\ell-1}}\big)Y_{\ell-1}^{(0)},$$

which is the Euler discretization of the solution of the SDE

$$dY_t = dM_t\, Y_t, \qquad Y_0 = W_0 x.$$

Let $\Phi_{s,t}$ denote the solution operator of this SDE, so that $Y_t = \Phi_{0,t}Y_0$. Hayou & Yang (2023) show that this Euler approximation converges at rate $O(L^{-1/2})$, with a constant independent of the width n. After taking the depth limit, the resulting finite-width SDE is an interacting particle system. They then compare each coordinate of this system with its McKean–Vlasov limit and obtain an $O(n^{-1/2})$ approximation error in $\mathbb{L}_2$. Together, these two estimates establish that the width and depth limits commute at initialization.

However, after one GD update, the residual blocks are no longer independent across layers. Indeed,

$$A_\ell^{(1),\tau} = I + \frac{1}{\sqrt{L}}W_\ell^{(1),\tau} = A_\ell^{(0)} - \frac{\gamma_{n,L}^{(0),\tau}}{nL}B_\ell^{(0)}(Y_{\ell-1}^{(0)})^\top,$$

where $Y_{\ell-1}^{(0)} = A_{\ell-1}^{(0)} \cdots A_1^{(0)}Y_0, \quad B_\ell^{(0)} = (A_{\ell+1}^{(0)})^\top \cdots (A_L^{(0)})^\top \xi$ by (3).

The forward state $Y_{\ell-1}^{(0)}$ depends on $A_1^{(0)}, \dots, A_{\ell-1}^{(0)}$, while the backward state $B_\ell^{(0)}$ depends on $A_{\ell+1}^{(0)}, \dots, A_L^{(0)}$. Therefore, $W_\ell^{(1),\tau}$ depends on the initial weights for $j \neq \ell$, so the updated residual blocks are coupled across layers and the independent-increment Brownian coupling used at initialization is no longer available. This coupling also changes the structure of the subsequent width limit. At initialization, the forward state is simply $Y_n^{(0)}(t) = \Phi_{0,t}Y_0$. After one GD update, the discrete forward recursion is

$$Y_\ell^{(1),\tau} - Y_{\ell-1}^{(1),\tau} = \frac{1}{\sqrt{L}}W_\ell Y_{\ell-1}^{(1),\tau} - \frac{\gamma_{n,L}^{(0),\tau}}{nL}B_\ell^{(0)}(Y_{\ell-1}^{(0)})^\top Y_{\ell-1}^{(1),\tau}.$$

As $L \to \infty$, the Gaussian term converges to the matrix-Brownian dynamics, while the L^{-1} gradient term becomes a drift in the depth variable. The continuous-depth state therefore satisfies

$$\mathrm{d}Y_n^{(1),\tau}(t) = \mathrm{d}M_t\, Y_n^{(1),\tau}(t) - \frac{\gamma_n^{(0),\tau}}{n}B_n^{(0)}(t)(Y_n^{(0)}(t))^\top Y_n^{(1),\tau}(t)\ \mathrm{d}t.$$

Integrating this SDE from 0 to t along the initialization flow Φ gives

$$Y_n^{(1),\tau}(t) = \Phi_{0,t}Y_0 - \gamma_n^{(0),\tau}\int_0^t \Phi_{s,t}B_n^{(0)}(s)\left\langle Y_n^{(0)}(s), Y_n^{(1),\tau}(s)\right\rangle_n \mathrm{d}s. \tag{5}$$

Therefore, proving the width convergence of $Y_n^{(1),\tau}$ also requires controlling the width limits of $B_n^{(0)}$, $\gamma_n^{(0),\tau}$, and $\langle Y_n^{(0)}, Y_n^{(1),\tau}\rangle_n$.

4.2 Proof Overview

To establish Theorem 4.1, we decompose the joint limit into two steps. First, at fixed width n, we compare the discrete network with a continuous-depth model whose output is denoted by $F_n^{(k),\tau}$, obtaining an $L^{-1/2}$ rate uniformly in n. We then compare $F_n^{(k),\tau}$ with its deterministic infinite-width limit $f_\infty^{(k),\tau}$ at rate $n^{-1/2}$:

$$f_{n,L}^{(k),\tau} \xrightarrow[L\to\infty]{L^{-1/2}} F_n^{(k),\tau} \xrightarrow[n\to\infty]{n^{-1/2}} f_\infty^{(k),\tau}.$$

Depth limit. For the discrete network, for $t \in I_\ell = (t_{\ell-1}, t_\ell]$, define

$$V_{n,L}^{(k)}(t) = \sum_{m<k} \frac{\gamma_{n,L}^{(m),\tau}}{n} B_{n,L}^{(m),\tau}(t_\ell)\big(Y_{n,L}^{(m),\tau}(t_{\ell-1})\big)^\top.$$

Let $Y_n^{(k),\tau}$ and $B_n^{(k),\tau}$ denote the finite-width continuous-depth forward and backward states, and let $\gamma_n^{(k),\tau}$ denote the clipped coefficient obtained in the continuous-depth model. Their recursive definitions are given in Appendix B.2. For the continuous-depth model, define

$$V_n^{(k)}(t) = \sum_{m<k} \frac{\gamma_n^{(m),\tau}}{n} B_n^{(m),\tau}(t)\big(Y_n^{(m),\tau}(t)\big)^\top.$$

Let $S_{i+1:j}^{n,L} := A_j^{(0)} \cdots A_{i+1}^{(0)}$, $\quad S_{j+1:j}^{n,L} = I$, denote the discrete initialization propagator. By construction, $S_{i+1:j}^{n,L}$ is the Euler discretization of Φ_{t_i,t_j}. At the grid points $t_j = j/L$, the discrete and continuous forward dynamics take the forms

$$Y_{n,L}^{(k),\tau}(t_j) = S_{1:j}^{n,L} Y_0 - \frac{1}{L}\sum_{\ell=1}^{j} S_{\ell+1:j}^{n,L} V_{n,L}^{(k)}(t_\ell) Y_{n,L}^{(k),\tau}(t_{\ell-1}),$$

$$Y_n^{(k),\tau}(t_j) = \Phi_{0,t_j} Y_0 - \int_0^{t_j} \Phi_{s,t_j} V_n^{(k)}(s) Y_n^{(k),\tau}(s)\, \mathrm{d}s.$$

For each fixed width n, the standard Euler–Maruyama estimate (Kloeden & Pearson, 1977) gives $\big\|S_{1:j}^{n,L} Y_0 - \Phi_{0,t_j} Y_0\big\|_{\mathbb{L}_2} = O(L^{-1/2})$, with a constant that depends on n. This is insufficient for the trained recursion, where $S_{\ell+1:j}^{n,L}$ and Φ_{s,t_j} act on $V_{n,L}^{(k)} Y_{n,L}^{(k),\tau}$ and $V_n^{(k)} Y_n^{(k),\tau}$ generated at each training iteration. We therefore require an operator-norm estimate that is uniform in the width. In Appendix B.1, we remove this width dependence and show that

$$\sup_{n\ge 1} \max_{0\le i\le j\le L} \Big\| \big\| S_{i+1:j}^{n,L} - \Phi_{t_i,t_j} \big\|_{\mathrm{op}} \Big\|_{\mathbb{L}_2} \le C_{\mathrm{Eul}} L^{-1/2},$$

where C_{Eul} is independent of both n and L (see Proposition B.2).

We then show by induction that this rate is preserved for any finite number of GD updates, following the proof chain below

$$\begin{array}{ccccccc}
\big(Y_{n,L}^{(k),\tau}, B_{n,L}^{(k),\tau}\big) & \longrightarrow & \gamma_{n,L}^{(k),\tau} & \longrightarrow & V_{n,L}^{(k+1)} & \longrightarrow & \big(Y_{n,L}^{(k+1),\tau}, B_{n,L}^{(k+1),\tau}\big) \\
\downarrow O(L^{-1/2}) & & \downarrow O(L^{-1/2}) & & \downarrow O(L^{-1/2}) & & \downarrow O(L^{-1/2}) \\
\big(Y_n^{(k),\tau}, B_n^{(k),\tau}\big) & \longrightarrow & \gamma_n^{(k),\tau} & \longrightarrow & V_n^{(k+1)} & \longrightarrow & \big(Y_n^{(k+1),\tau}, B_n^{(k+1),\tau}\big)
\end{array}$$

Clipping yields $\|V_{n,L}^{(k)}\|_{L_t^2(\mathrm{F})} \vee \|V_n^{(k)}\|_{L_t^2(\mathrm{F})} \le k\eta\tau$. Since T is fixed, this bound is uniform over n and L for all $k \le T$. The $O(L^{-1/2})$ rate is therefore preserved throughout the training induction. This gives the following width-uniform depth comparison.

Proposition 4.3 (Width-uniform depth discretization)**.** For every fixed integer $T \geq 0$, there exists $C_{\mathrm{D}} < \infty$, independent of n and L, such that

$$\max_{0 \leq k \leq T} \left\| f_{n,L}^{(k),\tau} - F_n^{(k),\tau} \right\|_{\mathbb{L}_2} \leq C_{\mathrm{D}} L^{-1/2}.$$

The complete proof is given in Appendix B.2.

Width limit. The one-step expression (5) already shows the structure that persists at later training iterations. At iteration k, the forward state involves $B_n^{(r),\tau}$, $\gamma_n^{(r),\tau}$, and $\langle Y_n^{(r),\tau}, Y_n^{(k),\tau} \rangle_n$ for $r < k$, in addition to the matrix-Brownian flow Φ. Since $Y_n^{(k),\tau} \in \mathbb{R}^n$, the state space itself changes with n, so there is no direct statewise comparison across widths. We therefore introduce an intermediate n-dimensional surrogate that keeps the finite-width matrix flow Φ, while replacing $\gamma_n^{(r),\tau}$ and the empirical inner products by their limiting values. Let $y^{(k)}$ and $b^{(k)}$ denote the limiting forward and backward states, and let γ_k^τ denote the limiting value of $\gamma_n^{(k),\tau}$ (see Appendix C). Define the surrogate states

$$\widetilde{Y}_n^{(k)}(t) = \Phi_{0,t} Y_0 - \sum_{r<k} \gamma_r^\tau \int_0^t \Phi_{s,t} \widetilde{B}_n^{(r)}(s) \left\langle y^{(r)}(s), y^{(k)}(s) \right\rangle \, \mathrm{d}s,$$

$$\widetilde{B}_n^{(k)}(t) = \Phi_{t,1}^\top \xi - \sum_{r<k} \gamma_r^\tau \int_t^1 \Phi_{t,s}^\top \widetilde{Y}_n^{(r)}(s) \left\langle b^{(r)}(s), b^{(k)}(s) \right\rangle \, \mathrm{d}s.$$

Recall that the continuous-depth output is $F_n^{(k),\tau} := \langle \xi, Y_n^{(k),\tau}(1) \rangle_n$. The surrogate yields the decomposition

$$F_n^{(k),\tau} - f_\infty^{(k),\tau} = \underbrace{\left\langle \xi, Y_n^{(k),\tau}(1) - \widetilde{Y}_n^{(k)}(1) \right\rangle_n}_{\text{(I)}} + \underbrace{\left(\left\langle \xi, \widetilde{Y}_n^{(k)}(1) \right\rangle_n - f_\infty^{(k),\tau} \right)}_{\text{(II)}}.$$

Term (I) measures the discrepancy between the finite-width model and the surrogate, while Term (II) compares the surrogate output with its deterministic limit. To control (I), we compare the finite-width and surrogate states inductively through the following chain:

$$\begin{array}{ccccc}
(Y_n^{(k),\tau}, B_n^{(k),\tau}) & \longrightarrow & \gamma_n^{(k),\tau} & \longrightarrow & (Y_n^{(k+1),\tau}, B_n^{(k+1),\tau}) \\
\downarrow O(n^{-1/2}) & & \downarrow O(n^{-1/2}) & & \downarrow O(n^{-1/2}) \\
(\widetilde{Y}_n^{(k)}, \widetilde{B}_n^{(k)}) & \longrightarrow & \gamma_k^\tau & \longrightarrow & (\widetilde{Y}_n^{(k+1)}, \widetilde{B}_n^{(k+1)})
\end{array}$$

This gives $\text{(I)} = O(n^{-1/2})$ for every fixed k. For (II), the surrogate output approaches its deterministic limit in two steps:

$$\left\langle \xi, \widetilde{Y}_n^{(k)}(1) \right\rangle_n \xrightarrow{n^{-1/2}} \mathbb{E}\left[\left\langle \xi, \widetilde{Y}_n^{(k)}(1) \right\rangle_n \middle| \Phi \right] \xrightarrow{n^{-1}} f_\infty^{(k),\tau}.$$

The first step conditions on the matrix flow Φ and averages out the Gaussian randomness in Y_0 and ξ, reducing the conditional mean to normalized traces of products of Φ. The second step shows that these finite-width traces converge to their infinite-width limits at rate $O(n^{-1})$, yielding $f_\infty^{(k),\tau}$.

Combining the bounds for (I) and (II) gives the following proposition.

Proposition 4.4 (Width limit of the continuous-depth model)**.** For every fixed integer $T \geq 0$, there exists $C_{\mathrm{W}} < \infty$, independent of n, such that

$$\max_{0 \leq k \leq T} \left\| F_n^{(k),\tau} - f_\infty^{(k),\tau} \right\|_{\mathbb{L}_2} \leq C_{\mathrm{W}} n^{-1/2}.$$

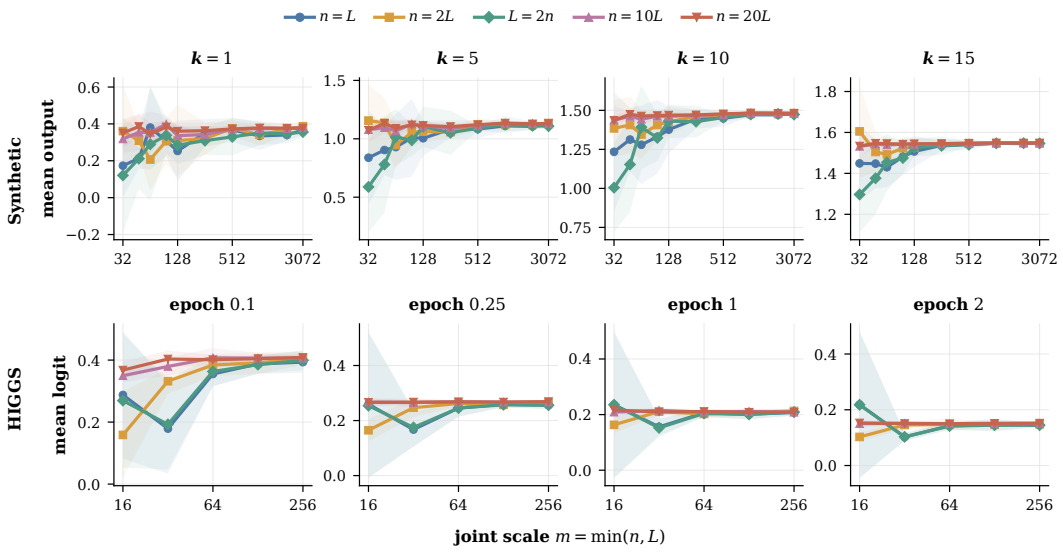


Figure 2: **Commutativity of the width and depth limits during training.** Each curve corresponds to a fixed width–depth scaling path and is plotted against the joint scale $m = \min(n, L)$. The top row shows the synthetic regression output at training iterations $k = 1, 5, 10, 15$, and the bottom row shows the HIGGS logit margin at four training checkpoints. For each task, all configurations are evaluated at the same fixed input. Curves show the mean across 5 model seeds, with shaded regions denoting ± 1 standard deviation.

The proof is given in Appendix C.

Propositions 4.3 and 4.4 control the depth and width errors, respectively. By the triangle inequality,

$$\max_{0 \le k \le T} \|f_{n,L}^{(k),\tau} - f_{\infty}^{(k),\tau}\|_{\mathbb{L}_2} \le C_{\mathrm{D}} L^{-1/2} + C_{\mathrm{W}} n^{-1/2},$$

which proves Theorem 4.1.

5 Experiments

In this section, we empirically study the commutativity of the width and depth limits during training and characterize their finite-size convergence rates. Additional experiments in Appendix G include a single-datum setting aligned with our theory and an empirical nonlinear extension across several activation functions.

5.1 Commutativity during Training

We evaluate joint width–depth scaling on synthetic regression and HIGGS binary classification.

Synthetic regression. We generate $N = 128$ samples $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ in dimension $d = 16$, where $\Sigma = Q \Lambda Q^\top$ has a geometrically spaced spectrum with condition number 10. For fixed orthonormal directions θ_1, θ_2, the regression targets are

$$\widetilde{y}_i = \theta_1^\top x_i + 0.5\big((\theta_2^\top x_i)^2 - 1\big), \qquad y_i = \widetilde{y}_i \left(\frac{1}{N} \sum_{j=1}^{N} \widetilde{y}_j^2 \right)^{-1/2}.$$

We train for 15 full-batch GD iterations with quadratic loss.

HIGGS (Baldi et al., 2014). HIGGS is a binary classification benchmark with 28 input features. We train the class-1 versus class-0 logit margin using binary logistic loss and minibatch SGD for two epochs with batch size 4096.We use the same data split and minibatch sequence for every width–depth configuration.

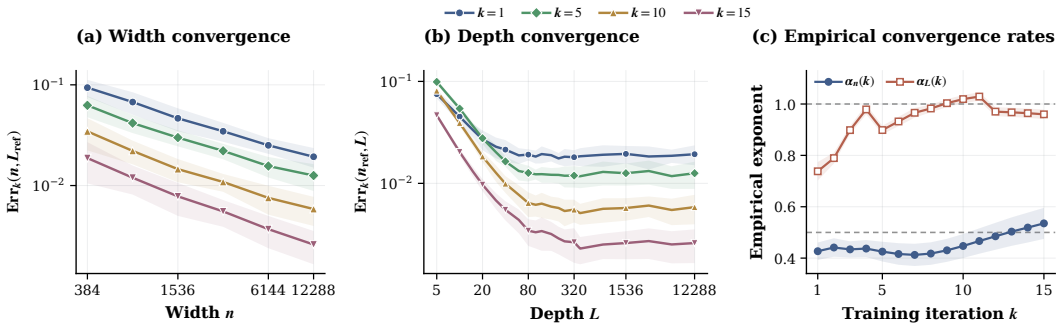


Figure 3: **Width–depth convergence rates.** (a) RMS output error as width n increases with depth fixed at $L = 12{,}288$. (b) RMS output error as depth L increases with width fixed at $n = 12{,}288$. Curves in (a)–(b) correspond to training iterations $k \in \{1, 5, 10, 15\}$. (c) Empirical convergence exponents for width and depth over training. The width rate remains close to $n^{-1/2}$, while the depth rate approaches an empirical L^{-1} rate.

For both tasks, we evaluate five width–depth scaling paths, $n = L$, $n = 2L$, $L = 2n$, $n = 10L$, and $n = 20L$, indexed by the common scale $m = \min(n, L)$. Along each path, both width and depth increase while their ratio remains fixed. As shown in Figure 2, the outputs differ across paths at small scales but approach a common value as m increases, consistent with the commutativity established by our theory.

5.2 Width–Depth Convergence Rate

We next characterize how the finite-size error scales with width and depth at different training iterations. Since the limiting output is not available in closed form, we construct an empirical reference $\widehat{f}_\infty^{(k)}$ by averaging the outputs at the largest joint scale $m = 3072$ over the five width–depth scaling paths and five seeds. We measure the RMS output error over the $N = 128$ samples,

$$\mathrm{Err}_k(n, L) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(f_{n,L}^{(k)}(x_i) - \widehat{f}_\infty^{(k)}(x_i) \right)^2 \right)^{1/2}.$$

To quantify how the finite-size error depends on width and depth, we vary one dimension while fixing the other at 12,288. Figure 3(a)–(b) shows the sweeps at $k \in \{1, 5, 10, 15\}$. The width error decreases steadily with n, whereas the depth error drops rapidly at small L and then plateaus as finite-width effects become dominant.

Let $n_{\mathrm{ref}} = L_{\mathrm{ref}} = 12{,}288$, and let $\mathrm{Err}_k^{\mathrm{floor}}$ denote the plateau level of the depth sweep at iteration k. To quantify the convergence rates, we fit

$$\mathrm{Err}_k(n, L_{\mathrm{ref}}) \propto n^{-\alpha_n(k)}, \qquad \mathrm{Err}_k(n_{\mathrm{ref}}, L) \propto L^{-\alpha_L(k)},$$

where the depth exponent is estimated over the decreasing regime of the RMS error curve. Figure 3(c) shows the fitted exponents over training. The width rate remains close to $n^{-1/2}$ throughout training, whereas the depth-dependent error decays more rapidly and approaches an empirical L^{-1} after the initial iterations.

Remark 5.1 (Depth rate)**.** The $L^{-1/2}$ rate in our Theorem 4.1 is an upper bound induced by the discrete-to-continuous estimates in the proof, and is not claimed to be tight. The faster empirical depth rate suggests that the finite-depth bound may be sharpened.

6 Conclusion

We studied whether width and depth limits continue to commute during training in linear residual networks. Under ordinary gradient descent, the limits commute in probability after any fixed number of training iterations, while the depth-first $\mathbb{L}_2$ limit already fails after one step at every fixed finite width. Under clipped gradient descent, we prove commutativity in $\mathbb{L}_2$ for any fixed number of training iterations, with errors of order $n^{-1/2}$ and $L^{-1/2}$.

AI USE STATEMENT

We used OpenAI Codex to assist with proof auditing, consistency checking, manuscript revision, and LaTeX formatting. We reviewed all AI-assisted work and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

All theoretical results are proved in the appendices, with the model, initialization, training dynamics, and limiting objects specified in the main text. The empirical results use synthetic data and the publicly available HIGGS dataset. Experimental settings, hyperparameters, and implementation details are provided in Appendix G.

REFERENCES

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, February 2011. ISSN 1557-9654. doi: 10.1109/tit.2010.2094817. URL http://dx.doi.org/10.1109/TIT.2010.2094817.

Philippe Biane and Roland Speicher. Stochastic calculus with respect to free brownian motion and analysis on wigner space. *Probability theory and related fields*, 112(3):373–409, 1998.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit, 2023. URL https://arxiv.org/abs/2309.16620.

Louis-Pierre Chaintron, Lénaïc Chizat, and Javier Maass. Resnets of all shapes and sizes: Convergence of training dynamics in the large-scale limit, 2026. URL https://arxiv.org/abs/2603.18168.

Lénaïc Chizat. The hidden width of deep resnets: Tight error bounds and phase diagram, 2026. URL https://arxiv.org/abs/2509.10167.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2018.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2019.

Soufiane Hayou. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2022.

Soufiane Hayou. Commutative width and depth scaling in deep neural networks, 2023. URL https://arxiv.org/abs/2310.01683.

Soufiane Hayou. A proof of learning rate transfer under μP. *arXiv preprint arXiv:2511.01734*, 2025.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning*, pp. 12700–12723. PMLR, 2023.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1324–1332. PMLR, 13–15 Apr 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Peter E Kloeden and RA Pearson. The numerical solution of stochastic differential equations. *The ANZIAM Journal*, 20(1):8–12, 1977.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Mufan Li, Mihai Nica, and Dan Roy. The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *Advances in Neural Information Processing Systems*, 35:10795–10808, 2022.

Pierre Marion, Adeline Fermanian, Gérard Biau, and Jean-Philippe Vert. Scaling resnets in the large-depth regime. *arXiv*, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281, 2023.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1126–1136. PMLR, 2020.

Catherine Emily Iska Redelmeier. Real second-order freeness and the asymptotic real second-order freeness of several real matrix models. *International Mathematics Research Notices*, 2014(12): 3353–3395, 2014.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

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*, volume 2024, pp. 55099–55150, 2024.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213. IEEE, 2022.

A Proof of the $\mathbb{L}_2$ Divergence at $k = 1$

This section proves Proposition 3.2.

Proof. At initialization, write $A_\ell^{(0)} = I + L^{-1/2} W_\ell$, $Y_\ell := Y_\ell^{(0)}$, $B_\ell := B_\ell^{(0)}$, and $f_{n,L} := f_{n,L}^{(0)}$.

Set $u_n := 1/(4n)$ and $b_n := 1 + u_n$. Define $E_{n,L}$ as the event on which

(i) $1 \le (Y_0)_i \le 2$ and $1 \le \xi_i \le 2$ for every i;

(ii) $u_n\sqrt{L} \le (W_\ell)_{11} \le 2u_n\sqrt{L}$ for every $\ell = 1, \dots, L$;

(iii) $0 \le (W_\ell)_{ij} \le 1$ for every ℓ and $(i, j) \ne (1, 1)$.

Since $q_0 > 0$, condition (i) has strictly positive probability. On $E_{n,L}$, every entry of $A_\ell^{(0)}$ is nonnegative and $(A_\ell^{(0)})_{11} \geq b_n$ for all ℓ.

Hence, the forward recursion gives

$$(Y_\ell)_1 \geq b_n^\ell, \qquad (Y_\ell)_i \geq 1, \tag{6}$$

for every ℓ and i. Similarly, the backward recursion gives

$$(B_\ell)_1 \geq b_n^{L-\ell}, \qquad (B_\ell)_i \geq 1. \tag{7}$$

Therefore,

$$f_{n,L} = \frac{1}{n}\xi^\top Y_L \geq \frac{b_n^L}{n}.$$

Since $b_n > 1$, for all sufficiently large L,

$$\varepsilon_{n,L}^{(0)} = f_{n,L} - y_\star \geq \frac{b_n^L}{2n} > 0. \tag{8}$$

The first GD update is $(A_\ell^{(1)})_{ij} = (A_\ell^{(0)})_{ij} - \frac{\eta\varepsilon_{n,L}^{(0)}}{nL}(B_\ell)_i(Y_{\ell-1})_j$.

On $E_{n,L}$,

$$\frac{\eta\varepsilon_{n,L}^{(0)}}{nL}(B_\ell)_i(Y_{\ell-1})_j \geq \frac{\eta b_n^L}{2n^2L} \quad \text{for all } i,j,$$

while the first coordinate satisfies the stronger bound

$$\frac{\eta\varepsilon_{n,L}^{(0)}}{nL}(B_\ell)_1(Y_{\ell-1})_1 \geq \frac{\eta b_n^{2L-1}}{2n^2L}.$$

Since $\eta b_n^L/(2n^2L) \to \infty$, all entries of $A_\ell^{(1)}$ are negative for sufficiently large L. Moreover, the stronger bound on the first coordinate gives

$$\left|(A_\ell^{(1)})_{11}\right| \geq d_{n,L} := \frac{\eta b_n^{2L-1}}{4n^2L}, \qquad \ell = 1,\dots,L. \tag{9}$$

Set $P_\ell := -A_\ell^{(1)}$. On $E_{n,L}$, every entry of P_ℓ is positive. Therefore,

$$\left|f_{n,L}^{(1)}\right| = \frac{1}{n}\xi^\top P_L \cdots P_1 Y_0 \geq \frac{1}{n} d_{n,L}^L. \tag{10}$$

It remains to estimate the probability of $E_{n,L}$. Define

$$p_{n,L} := \mathbb{P}\Big(u_n\sqrt{L} \leq (W_\ell)_{11} \leq 2u_n\sqrt{L}\Big), \qquad q_n := \mathbb{P}(0 \leq (W_\ell)_{ij} \leq 1),$$

where $(i,j) \neq (1,1)$. These probabilities do not depend on ℓ, since all entries of W_ℓ are identically distributed. Also, let

$$r_n := \mathbb{P}(1 \leq (Y_0)_i \leq 2,\ 1 \leq \xi_i \leq 2 \text{ for all } i) > 0.$$

By independence,

$$\mathbb{P}(E_{n,L}) = r_n\, p_{n,L}^L\, q_n^{(n^2-1)L}. \tag{11}$$

Since $(W_\ell)_{11} \sim \mathcal{N}(0, n^{-1})$, its density is $\phi_n(w) = \sqrt{n/(2\pi)}\exp(-nw^2/2)$, which is decreasing on $[0,\infty)$. Therefore,

$$p_{n,L} \geq u_n\sqrt{L}\sqrt{\frac{n}{2\pi}}\exp(-2nu_n^2L). \tag{12}$$

Combining (11) and (12) gives

$$\log\mathbb{P}(E_{n,L}) \geq -2nu_n^2L^2 + \frac{1}{2}L\log L + c_nL + \log r_n, \tag{13}$$

where $c_n := (n^2-1)\log q_n + \log u_n + \frac{1}{2}\log(n/(2\pi))$ depends only on n.

Using (10),

$$\mathbb{E}\big|f_{n,L}^{(1)}\big|^2 \geq \mathbb{P}(E_{n,L})\,\frac{d_{n,L}^{2L}}{n^2}.$$

Together with (9) and (13), this yields

$$\log \mathbb{E}\big|f_{n,L}^{(1)}\big|^2 \geq \kappa_n L^2 - \frac{3}{2}L\log L + O_{n,\eta}(L), \tag{14}$$

where $\kappa_n := 4\log b_n - 2nu_n^2$.

Since $u_n = 1/(4n) \leq 1$ and $\log(1+x) \geq x/2$ for $x \in [0,1]$,

$$\kappa_n = 4\log(1+u_n) - 2nu_n^2 \geq 2u_n - 2nu_n^2 = \frac{3}{8n} > 0.$$

Dividing (14) by L^2 and letting $L \to \infty$ gives

$$\mathbb{E}\big|f_{n,L}^{(1)}\big|^2 \longrightarrow +\infty.$$

This proves Proposition 3.2. □

B Depth Discretization

We prove Proposition 4.3 by comparing the trained residual network with the continuous-depth finite-width model. We first compare the discrete and continuous representations $Y_{n,L}^{(k),\tau}$ and $Y_n^{(k),\tau}$. The corresponding estimate for $B_{n,L}^{(k),\tau}$ and $B_n^{(k),\tau}$ follows by depth reversal and transposition; see Appendix B.2. Here, $\mathcal{J}_n^{(k),\tau}$ denotes the continuous-depth analogue of $J_{n,L}^{(k),\tau}$, defined in Appendix B.2.

The depth comparison follows the chain

$$S_{i+1:j}^{n,L} \xrightarrow{L^{-1/2}} \Phi_{t_i,t_j},$$

followed by the finite-training induction

$$V_{n,L}^{(k)} \xrightarrow{L^{-1/2}} V_n^{(k)} \implies \big(Y_{n,L}^{(k),\tau}, B_{n,L}^{(k),\tau}\big) \xrightarrow{L^{-1/2}} \big(Y_n^{(k),\tau}, B_n^{(k),\tau}\big)$$
$$\implies \big(f_{n,L}^{(k),\tau}, J_{n,L}^{(k),\tau}\big) \xrightarrow{L^{-1/2}} \big(F_n^{(k),\tau}, \mathcal{J}_n^{(k),\tau}\big) \implies V_{n,L}^{(k+1)} \xrightarrow{L^{-1/2}} V_n^{(k+1)}.$$

B.1 Width-uniform base approximation

Let

$$t_i := \frac{i}{L}, \qquad S_{i+1:j}^{n,L} := A_j^{(0)} \cdots A_{i+1}^{(0)}, \qquad 0 \leq i \leq j \leq L,$$

with $S_{j+1:j}^{n,L} = I$. Couple the Gaussian residual matrices with an $n \times n$ matrix Brownian motion M through

$$M_{t_j} - M_{t_{j-1}} = L^{-1/2} W_j,$$

and let $\Phi_{s,t}$ solve

$$\mathrm{d}\Phi_{s,t} = \mathrm{d}M_t\,\Phi_{s,t}, \qquad \Phi_{s,s} = I.$$

Lemma B.1 (Maximal inequality for matrix-Brownian integrals)**.** Let $M_t = n^{-1/2}(\beta_{\alpha\beta}(t))_{\alpha,\beta=1}^n$, where the $\beta_{\alpha\beta}$ are independent standard Brownian motions. Let ψ be a predictable $\mathbb{R}^n$-valued process satisfying $\int_a^b \|\psi_s\|_{L^r(\Omega;\mathbb{R}^n)}^2\,\mathrm{d}s < \infty$ for some $r \geq 2$, and set $Z_t := \int_a^t \mathrm{d}M_s\,\psi_s$ for $a \leq t \leq b$. Then

$$\Big\| \sup_{a\leq u\leq t} \|Z_u\|_2 \Big\|_{L^r} \leq K_{r,n}\left(\int_a^t \|\psi_s\|_{L^r(\Omega;\mathbb{R}^n)}^2\,\mathrm{d}s\right)^{1/2}, \qquad K_{r,n} := \frac{r}{r-1}\sqrt{1+\frac{r-2}{n}}. \tag{15}$$

Proof. Writing $Z_t = (Z_1(t), \dots, Z_n(t))^\top$, we have $\mathrm{d}Z_\alpha(t) = n^{-1/2} \sum_{\beta=1}^n \psi_\beta(t)\, \mathrm{d}\beta_{\alpha\beta}(t)$ and hence $\mathrm{d}\langle Z_\alpha, Z_\gamma\rangle_t = \delta_{\alpha\gamma} n^{-1} \|\psi_t\|_2^2\, \mathrm{d}t$. Itô's formula applied to $\|Z_t\|_2^r$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|Z_t\|_2^r = \frac{r(r+n-2)}{2n}\mathbb{E}\big[\|Z_t\|_2^{r-2}\|\psi_t\|_2^2\big]. \tag{16}$$

By Hölder's inequality,

$$\mathbb{E}\big[\|Z_t\|_2^{r-2}\|\psi_t\|_2^2\big] \le \big(\mathbb{E}\|Z_t\|_2^r\big)^{(r-2)/r} \|\psi_t\|^2_{L^r(\Omega;\mathbb{R}^n)}.$$

Substituting this into (16) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\mathbb{E}\|Z_t\|_2^r\big)^{2/r} \le \frac{r+n-2}{n}\|\psi_t\|^2_{L^r(\Omega;\mathbb{R}^n)}.$$

Since $Z_a = 0$, integration gives

$$\|Z_t\|_{L^r(\Omega;\mathbb{R}^n)} \le \sqrt{1 + \frac{r-2}{n}} \left(\int_a^t \|\psi_s\|^2_{L^r(\Omega;\mathbb{R}^n)}\, \mathrm{d}s\right)^{1/2}. \tag{17}$$

Since $\|Z_t\|_2$ is a nonnegative submartingale, Doob's inequality and (17) give (15). □

Proposition B.2 (Width-uniform discrete-to-continuous comparison)**.** For every $p \ge 2$, there are finite constants $C_{\mathrm{Eul}}(p)$, $C_{\mathrm{base}}(p)$, and $C_{\mathrm{reg}}(p)$, independent of n and L, such that

$$\sup_{n\ge1} \max_{0\le i\le j\le L} \left\| \left\| S^{n,L}_{i+1:j} - \Phi_{t_i,t_j} \right\|_{\mathrm{op}} \right\|_{L^p} \le C_{\mathrm{Eul}}(p) L^{-1/2}, \tag{18}$$

$$\sup_{n\ge1} \sup_{0\le s\le1} \left\| \sup_{s\le t\le1} \|\Phi_{s,t}\|_{\mathrm{op}} \right\|_{L^p} + \sup_{n,L} \max_{0\le i\le L} \left\| \max_{i\le j\le L} \|S^{n,L}_{i+1:j}\|_{\mathrm{op}} \right\|_{L^p} \le C_{\mathrm{base}}(p). \tag{19}$$

Moreover, for every $0 \le a \le b \le 1$,

$$\sup_{n\ge1} \left\| \sup_{a\le s\le b} \big(\|\Phi_{a,s} - I\|_{\mathrm{op}} + \|\Phi_{s,b} - I\|_{\mathrm{op}}\big) \right\|_{L^p} \le C_{\mathrm{reg}}(p)\sqrt{b-a}. \tag{20}$$

Proof. Fix $0 \le i \le L$ and a deterministic Euclidean unit vector v, and set $U_t := \Phi_{t_i,t} v$ for $t \in [t_i, 1]$. Writing $M_t = n^{-1/2}(\beta_{\alpha\beta}(t))_{\alpha,\beta=1}^n$, where the $\beta_{\alpha\beta}$ are independent standard Brownian motions, U solves $\mathrm{d}U_t = \mathrm{d}M_t\, U_t$ with $U_{t_i} = v$.

We first check the assumptions of the standard strong Euler–Maruyama estimate. For $u, w \in \mathbb{R}^n$ and $\delta > 0$, independence of the Brownian entries gives

$$\mathbb{E}\left[\|(M_{t+\delta} - M_t)(u-w)\|_2^2\right] = \delta\|u-w\|_2^2.$$

Thus the diffusion coefficient is globally Lipschitz with constant one, uniformly in n. The same identity with $u-w$ replaced by u gives the linear-growth condition, again with a constant independent of n.

Under the coupling above, the Euler–Maruyama recursion on $\{t_j\}_{j=i}^L$ is

$$U^L_{t_j} = \big(I + L^{-1/2} W_j\big) U^L_{t_{j-1}}, \qquad U^L_{t_i} = v,$$

so $U^L_{t_j} = S^{n,L}_{i+1:j} v$. Therefore, for every fixed n and $p \ge 2$, the standard strong Euler–Maruyama theorem gives

$$\max_{i\le j\le L} \left\| S^{n,L}_{i+1:j} v - \Phi_{t_i,t_j} v \right\|_{L^p} \le C_{n,p} L^{-1/2}.$$

This estimate controls one deterministic direction, whereas (18) requires an operator-norm bound with a constant uniform in n. We therefore track the dependence on both the moment order and the width.

For every fixed n and $r \geq 2$, the standard moment estimate for globally Lipschitz SDEs gives $\int_{t_i}^1 \|U_s\|_{L^r}^2 \, \mathrm{d}s < \infty$. Taking $\psi_s = U_s$ in Lemma B.1 and using $U_t = v + \int_{t_i}^t \mathrm{d}M_s \, U_s$, set $F(t) := \| \sup_{t_i \leq u \leq t} \|U_u\|_2 \|_{L^r}$. Then

$$F(t)^2 \leq 2 + 2K_{r,n}^2 \int_{t_i}^t F(s)^2 \, \mathrm{d}s.$$

Gronwall's inequality gives

$$\left\| \sup_{t_i \leq t \leq 1} \|U_t\|_2 \right\|_{L^r} \leq \sqrt{2} \, e^{K_{r,n}^2}. \tag{21}$$

We next control the variation of U within one mesh interval. For $i < \ell \leq L$ and $s \in [t_{\ell-1}, t_\ell]$,

$$U_s - U_{t_{\ell-1}} = \int_{t_{\ell-1}}^s \mathrm{d}M_u \, U_u.$$

Lemma B.1 on $[t_{\ell-1}, t_\ell]$, together with (21), gives

$$\left\| \sup_{t_{\ell-1} \leq s \leq t_\ell} \|U_s - U_{t_{\ell-1}}\|_2 \right\|_{L^r} \leq \sqrt{2} \, K_{r,n} e^{K_{r,n}^2} L^{-1/2}. \tag{22}$$

Let U^L denote the continuous Euler interpolation,

$$U_t^L = v + \int_{t_i}^t \mathrm{d}M_s \, U_{\pi_L(s)}^L, \qquad \pi_L(s) = t_{\ell-1} \quad \text{for } s \in (t_{\ell-1}, t_\ell].$$

At the grid points, $U_{t_j}^L = S_{i+1:j}^{n,L} v$, and

$$U_t^L - U_t = \int_{t_i}^t \mathrm{d}M_s \left(U_{\pi_L(s)}^L - U_{\pi_L(s)} + U_{\pi_L(s)} - U_s \right).$$

The Euler interpolation U^L has finite r-th moments for every fixed n, L, and $r \geq 2$. Setting $E(t) := \| \sup_{t_i \leq u \leq t} \|U_u^L - U_u\|_2 \|_{L^r}$, Lemma B.1 and (22) give

$$E(t)^2 \leq 2K_{r,n}^2 \int_{t_i}^t E(s)^2 \, \mathrm{d}s + 4K_{r,n}^4 e^{2K_{r,n}^2} L^{-1}.$$

Gronwall's inequality therefore gives

$$\left\| \sup_{t_i \leq t \leq 1} \|U_t^L - U_t\|_2 \right\|_{L^r} \leq 2K_{r,n}^2 e^{2K_{r,n}^2} L^{-1/2}. \tag{23}$$

To obtain the operator-norm estimate uniformly in n, let $\mathcal{N}$ be a deterministic $1/2$-net of the Euclidean unit sphere with $|\mathcal{N}| \leq 5^n$, and set $r = 2\lceil \max\{n, p, 4\}/2 \rceil$. Then $r \geq n, p$, $5^{n/r} \leq 5$, and

$$K_{r,n}^2 = \left(\frac{r}{r-1} \right)^2 \left(1 + \frac{r-2}{n} \right) \leq 2(p+3). \tag{24}$$

The $1/2$-net inequality gives

$$\left\| S_{i+1:j}^{n,L} - \Phi_{t_i,t_j} \right\|_{\mathrm{op}} \leq 2 \max_{v \in \mathcal{N}} \left\| (S_{i+1:j}^{n,L} - \Phi_{t_i,t_j}) v \right\|_2.$$

Hence

$$\left\| \left\| S_{i+1:j}^{n,L} - \Phi_{t_i,t_j} \right\|_{\mathrm{op}} \right\|_{L^p} \leq 2|\mathcal{N}|^{1/r} \max_{v \in \mathcal{N}} \left\| \left\| (S_{i+1:j}^{n,L} - \Phi_{t_i,t_j}) v \right\|_2 \right\|_{L^r}.$$

Since $r \geq n$, the factor $|\mathcal{N}|^{1/r}$ is bounded uniformly in n, while (24) gives a bound on $K_{r,n}$ that is also independent of n. Applying (23), uniformly over the starting index i, proves (18).

Applying the same net estimate to (21), and starting the same SDE at an arbitrary deterministic $s \in [0,1]$, gives

$$\sup_{n\geq 1}\sup_{0\leq s\leq 1}\left\|\sup_{s\leq t\leq 1}\|\Phi_{s,t}\|_{\mathrm{op}}\right\|_{L^p} \leq 10\sqrt{2}\,e^{2(p+3)}.$$

Moreover, since $\|U_t^L\|_2 \leq \|U_t\|_2 + \|U_t^L - U_t\|_2$, (21) and (23), followed by the same net estimate, give

$$\sup_{n,L}\max_{0\leq i\leq L}\left\|\max_{i\leq j\leq L}\|S^{n,L}_{i+1:j}\|_{\mathrm{op}}\right\|_{L^p} \leq 10\sqrt{2}\,e^{2(p+3)} + 40(p+3)e^{4(p+3)}.$$

Thus, (19) holds with $C_{\mathrm{base}}(p) := 20\sqrt{2}\,e^{2(p+3)} + 40(p+3)e^{4(p+3)}$.

We also need the variation of the base flow over a short depth interval. Fix $0 \leq a \leq b \leq 1$. For every deterministic unit vector v,

$$(\Phi_{a,s} - I)v = \int_a^s \mathrm{d}M_u\,\Phi_{a,u}v, \qquad a \leq s \leq b.$$

Applying Lemma B.1 on $[a,b]$, using (21), and then applying the same net estimate gives

$$\sup_{n\geq 1}\left\|\sup_{a\leq s\leq b}\|\Phi_{a,s} - I\|_{\mathrm{op}}\right\|_{L^p} \leq C_{\mathrm{reg}}(p)\sqrt{b-a}.$$

It remains to control the flow when the right endpoint is fixed. For fixed b, define

$$\check{M}_u := (M_b - M_{b-u})^\top, \qquad 0 \leq u \leq b-a.$$

Then $\check{M}$ has the same law as M, and its forward flow satisfies $\check{\Phi}_{0,u} = \Phi_{b-u,b}^\top$.

Applying the preceding estimate to $\check{\Phi}$ and using $\|\Phi_{b-u,b}^\top - I\|_{\mathrm{op}} = \|\Phi_{b-u,b} - I\|_{\mathrm{op}}$ gives

$$\sup_{n\geq 1}\left\|\sup_{a\leq s\leq b}\|\Phi_{s,b} - I\|_{\mathrm{op}}\right\|_{L^p} \leq C_{\mathrm{reg}}(p)\sqrt{b-a}.$$

Combining the two estimates proves (20). □

B.2 Finite-training depth comparison

We now show that the $L^{-1/2}$ rate is preserved through every fixed training step.

We first define the finite-width continuous-depth training dynamics. Set $V_n^{(0)}(t) := 0$. Given $V_n^{(k)}$, let the forward and backward states satisfy

$$Y_n^{(k),\tau}(t) = \Phi_{0,t}Y_0 - \int_0^t \Phi_{s,t}V_n^{(k)}(s)Y_n^{(k),\tau}(s)\,\mathrm{d}s, \tag{25}$$

$$B_n^{(k),\tau}(t) = \Phi_{t,1}^\top\xi - \int_t^1 \Phi_{t,s}^\top V_n^{(k)}(s)^\top B_n^{(k),\tau}(s)\,\mathrm{d}s. \tag{26}$$

Let $F_n^{(k),\tau} := \langle \xi, Y_n^{(k),\tau}(1)\rangle_n$ and

$$\mathcal{J}_n^{(k),\tau} := \int_0^1 \|Y_n^{(k),\tau}(t)\|_n^2\|B_n^{(k),\tau}(t)\|_n^2\,\mathrm{d}t.$$

Define $\gamma_n^{(k),\tau} := \eta\big(F_n^{(k),\tau} - y_\star\big)\min\left\{1, \frac{\tau}{|F_n^{(k),\tau}-y_\star|\sqrt{\mathcal{J}_n^{(k),\tau}}}\right\}$, where the multiplier inside the minimum is set to one if the denominator vanishes. For $k < T$, set

$$V_n^{(k+1)}(t) = V_n^{(k)}(t) + \frac{\gamma_n^{(k),\tau}}{n}B_n^{(k),\tau}(t)\big(Y_n^{(k),\tau}(t)\big)^\top.$$

For $t \in I_\ell := (t_{\ell-1}, t_\ell]$, define the discrete cumulative update by

$$V_{n,L}^{(k)}(t) = \sum_{m<k}\frac{\gamma_{n,L}^{(m),\tau}}{n}B_{n,L}^{(m),\tau}(t_\ell)\big(Y_{n,L}^{(m),\tau}(t_{\ell-1})\big)^\top.$$

Iterating the continuous update gives

$$V_n^{(k)}(t) = \sum_{m<k} \frac{\gamma_n^{(m),\tau}}{n} B_n^{(m),\tau}(t)\big(Y_n^{(m),\tau}(t)\big)^\top.$$

For the cumulative updates above, we write $\|V\|_{L_t^2(\mathrm{F})} := \left(\int_0^1 \|V(t)\|_{\mathrm{F}}^2 \, \mathrm{d}t\right)^{1/2}$.

By the rank-one structure of the updates and the definitions of $J_{n,L}^{(m),\tau}$ and $\mathcal{J}_n^{(m),\tau}$,

$$\left(\frac{1}{L}\sum_{\ell=1}^{L} \left\| \frac{1}{n} B_{n,L}^{(m),\tau}(t_\ell)\big(Y_{n,L}^{(m),\tau}(t_{\ell-1})\big)^\top \right\|_{\mathrm{F}}^2\right)^{1/2} = \sqrt{J_{n,L}^{(m),\tau}},$$
$$\left(\int_0^1 \left\| \frac{1}{n} B_n^{(m),\tau}(t)\big(Y_n^{(m),\tau}(t)\big)^\top \right\|_{\mathrm{F}}^2 \, \mathrm{d}t\right)^{1/2} = \sqrt{\mathcal{J}_n^{(m),\tau}}.$$

By the definition of the clipped coefficients,

$$|\gamma_{n,L}^{(m),\tau}|\sqrt{J_{n,L}^{(m),\tau}} \le \eta\tau, \qquad |\gamma_n^{(m),\tau}|\sqrt{\mathcal{J}_n^{(m),\tau}} \le \eta\tau.$$

Summing over $m < k$ gives

$$\|V_{n,L}^{(k)}\|_{L_t^2(\mathrm{F})} \vee \|V_n^{(k)}\|_{L_t^2(\mathrm{F})} \le k\eta\tau \qquad \text{a.s.} \tag{27}$$

After reversing the depth variable and transposing the matrices, the backward equations take the same form as the forward equations. Indeed, if $B_{\ell-1} = A_\ell^\top B_\ell$, then

$$B_{L-i} = A_{L-i+1}^\top B_{L-i+1}, \qquad 1 \le i \le L,$$

which is a forward recursion in the reversed layer index. At initialization,

$$A_{L-i+1}^{(0)\top} = I + L^{-1/2} W_{L-i+1}^\top,$$

and $(W_L^\top, \dots, W_1^\top)$ has the same joint law as $(W_1, \dots, W_L)$.

For the continuous equation,

$$B(t) = \Phi_{t,1}^\top \xi - \int_t^1 \Phi_{t,s}^\top V(s)^\top B(s) \, \mathrm{d}s,$$

the substitutions $u = 1 - t$ and $r = 1 - s$ give

$$B(1-u) = \Phi_{1-u,1}^\top \xi - \int_0^u \Phi_{1-u,1-r}^\top V(1-r)^\top B(1-r) \, \mathrm{d}r. \tag{28}$$

The process $u \mapsto (M_1 - M_{1-u})^\top$ has the same law as M, and its forward flow from r to u is $\Phi_{1-u,1-r}^\top$. Moreover,

$$\int_0^1 \|V(1-r)^\top\|_{\mathrm{F}}^2 \, \mathrm{d}r = \int_0^1 \|V(s)\|_{\mathrm{F}}^2 \, \mathrm{d}s.$$

Hence the reversed backward equation has the same form as the forward equation, with the same Brownian law and the same $L_t^2(\mathrm{F})$ bound on the cumulative update. The estimates below therefore also hold for the backward equation. We prove only the forward estimate in detail.

Proof of Proposition 4.3. At $k = 0$, $V_{n,L}^{(0)} = V_n^{(0)} = 0$. We prove by induction that, for every finite $q \ge 2$,

$$\sup_{n \ge 1} \left\| \|V_{n,L}^{(k)} - V_n^{(k)}\|_{L_t^2(\mathrm{F})} \right\|_{L^q} \le C_V(k,q) L^{-1/2}. \tag{29}$$

Proposition F.1 and (27) give the moment bounds for the trained forward states used below.

For $s \in I_\ell$,

$$Y_n^{(k),\tau}(s) - Y_n^{(k),\tau}(t_{\ell-1}) = (\Phi_{t_{\ell-1},s} - I)Y_n^{(k),\tau}(t_{\ell-1}) - \int_{t_{\ell-1}}^{s} \Phi_{u,s} V_n^{(k)}(u) Y_n^{(k),\tau}(u)\, \mathrm{d}u.$$

Using (20), (27), and the moment bounds from Proposition F.1, Cauchy–Schwarz on each I_ℓ gives

$$\sup_{n\geq 1} \left\| \left(\sum_{\ell=1}^{L} \int_{I_\ell} \|Y_n^{(k),\tau}(s) - Y_n^{(k),\tau}(t_{\ell-1})\|_n^2 \, \mathrm{d}s \right)^{1/2} \right\|_{L^q} \leq C_{\mathrm{cell}}(k,q) L^{-1/2}. \tag{30}$$

Set $r := 4q$. At the grid points,

$$Y_{n,L}^{(k),\tau}(t_j) = S_{1:j}^{n,L} Y_0 - \frac{1}{L} \sum_{\ell=1}^{j} S_{\ell+1:j}^{n,L} V_{n,L}^{(k)}(t_\ell) Y_{n,L}^{(k),\tau}(t_{\ell-1}),$$

$$Y_n^{(k),\tau}(t_j) = \Phi_{0,t_j} Y_0 - \int_0^{t_j} \Phi_{s,t_j} V_n^{(k)}(s) Y_n^{(k),\tau}(s)\, \mathrm{d}s.$$

Since $V_{n,L}^{(k)}$ is constant on each I_ℓ, subtracting the two equations gives

$$\begin{aligned} & S_{\ell+1:j}^{n,L} V_{n,L}^{(k)}(t_\ell) Y_{n,L}^{(k),\tau}(t_{\ell-1}) - \Phi_{s,t_j} V_n^{(k)}(s) Y_n^{(k),\tau}(s) \\ = \; & \underbrace{\big(S_{\ell+1:j}^{n,L} - \Phi_{s,t_j}\big) V_{n,L}^{(k)}(t_\ell) Y_{n,L}^{(k),\tau}(t_{\ell-1})}_{\text{(a)}} + \underbrace{\Phi_{s,t_j}\big(V_{n,L}^{(k)}(t_\ell) - V_n^{(k)}(s)\big) Y_{n,L}^{(k),\tau}(t_{\ell-1})}_{\text{(b)}} \\ & + \underbrace{\Phi_{s,t_j} V_n^{(k)}(s)\big(Y_{n,L}^{(k),\tau}(t_{\ell-1}) - Y_n^{(k),\tau}(s)\big)}_{\text{(c)}}. \end{aligned}$$

For (a),

$$S_{\ell+1:j}^{n,L} - \Phi_{s,t_j} = \big(S_{\ell+1:j}^{n,L} - \Phi_{t_\ell,t_j}\big) + \Phi_{t_\ell,t_j}\big(I - \Phi_{s,t_\ell}\big).$$

Proposition B.2, (27), and the moment bounds for the forward states give

$$\left\| \max_{j\leq L} \left\| \sum_{\ell=1}^{j} \int_{I_\ell} \text{(a)}\, \mathrm{d}s \right\|_n \right\|_{L^{2r}} \leq C_a(k,r) L^{-1/2}.$$

For (b), Cauchy–Schwarz in depth gives

$$\begin{aligned} \max_{j\leq L} \left\| \sum_{\ell=1}^{j} \int_{I_\ell} \text{(b)}\, \mathrm{d}s \right\|_n \leq & \sup_{s\leq t} \|\Phi_{s,t}\|_{\mathrm{op}} \max_{\ell} \|Y_{n,L}^{(k),\tau}(t_{\ell-1})\|_n \\ & \times \|V_{n,L}^{(k)} - V_n^{(k)}\|_{L_t^2(\mathrm{F})}. \end{aligned}$$

Hölder's inequality, (19), the state moment bounds, and the induction hypothesis at moment $4r$ therefore give

$$\left\| \max_{j\leq L} \left\| \sum_{\ell=1}^{j} \int_{I_\ell} \text{(b)}\, \mathrm{d}s \right\|_n \right\|_{L^{2r}} \leq C_b(k,r) L^{-1/2}.$$

For (c), write

$$Y_{n,L}^{(k),\tau}(t_{\ell-1}) - Y_n^{(k),\tau}(s) = E_{\ell-1}^{(k)} + Y_n^{(k),\tau}(t_{\ell-1}) - Y_n^{(k),\tau}(s),$$

where $E_{\ell-1}^{(k)} := Y_{n,L}^{(k),\tau}(t_{\ell-1}) - Y_n^{(k),\tau}(t_{\ell-1})$. The difference $Y_n^{(k),\tau}(t_{\ell-1}) - Y_n^{(k),\tau}(s)$ is controlled by (30), while $E_{\ell-1}^{(k)}$ remains in the linear equation for the current grid-point difference. Proposition B.2 also gives an L^{2r} bound of order $L^{-1/2}$ for $(S_{1:j}^{n,L} - \Phi_{0,t_j})Y_0$.

Consequently,

$$E_j^{(k)} = R_j^{(k)} - \sum_{\ell=1}^{j} \int_{I_\ell} \Phi_{s,t_j} V_n^{(k)}(s) E_{\ell-1}^{(k)} \, \mathrm{d}s,$$

where $E_j^{(k)} := Y_{n,L}^{(k),\tau}(t_j) - Y_n^{(k),\tau}(t_j)$ and $\| \max_j \|R_j^{(k)}\|_n \|_{L^{2r}} \le C_R(k,r) L^{-1/2}$. Proposition F.2 with $\Lambda = k\eta\tau$ therefore gives

$$\sup_{n\ge 1} \left\| \max_{0\le j\le L} \|Y_{n,L}^{(k),\tau}(t_j) - Y_n^{(k),\tau}(t_j)\|_n \right\|_{L^r} \le C_Y(k,r) L^{-1/2}. \tag{31}$$

The reversal and transposition argument above gives the corresponding grid-point bound for the backward states and, for $s \in I_\ell$, the bound for $B_n^{(k),\tau}(s) - B_n^{(k),\tau}(t_\ell)$:

$$\sup_{n\ge 1} \left\| \max_{0\le j\le L} \|B_{n,L}^{(k),\tau}(t_j) - B_n^{(k),\tau}(t_j)\|_n \right\|_{L^r} \le C_B(k,r) L^{-1/2},$$

and

$$\sup_{n\ge 1} \left\| \left(\sum_{\ell=1}^{L} \int_{I_\ell} \|B_n^{(k),\tau}(s) - B_n^{(k),\tau}(t_\ell)\|_n^2 \, \mathrm{d}s \right)^{1/2} \right\|_{L^r} \le C_{B,\mathrm{reg}}(k,r) L^{-1/2}.$$

It remains to close the induction. The output identity

$$f_{n,L}^{(k),\tau} - F_n^{(k),\tau} = \left\langle \xi, Y_{n,L}^{(k),\tau}(1) - Y_n^{(k),\tau}(1) \right\rangle_n$$

together with the uniform moment bound for $\|\xi\|_n$ and (31) gives

$$\sup_n \|f_{n,L}^{(k),\tau} - F_n^{(k),\tau}\|_{L^q} \le C_f(k,q) L^{-1/2}.$$

The forward and backward state bounds, together with (30) and the corresponding bound for $B_n^{(k),\tau}$, give

$$\sup_n \|J_{n,L}^{(k),\tau} - \mathcal{J}_n^{(k),\tau}\|_{L^q} \le C_J(k,q) L^{-1/2},$$

using $|a^2b^2 - c^2d^2| \le (a+c)|a-c|b^2 + c^2(b+d)|b-d|$.

If $k < T$, define

$$G_{n,L}^{(k)}(t) = \frac{f_{n,L}^{(k),\tau} - y_\star}{n} B_{n,L}^{(k),\tau}(t_\ell) \big(Y_{n,L}^{(k),\tau}(t_{\ell-1})\big)^\top, \qquad t \in I_\ell,$$

and

$$G_n^{(k)}(t) = \frac{F_n^{(k),\tau} - y_\star}{n} B_n^{(k),\tau}(t) \big(Y_n^{(k),\tau}(t)\big)^\top.$$

Since

$$\left\| \frac{1}{n} B_{n,L}^{(k),\tau}(t_\ell) \big(Y_{n,L}^{(k),\tau}(t_{\ell-1})\big)^\top \right\|_{\mathrm{F}} = \|B_{n,L}^{(k),\tau}(t_\ell)\|_n \|Y_{n,L}^{(k),\tau}(t_{\ell-1})\|_n,$$

$$\left\| \frac{1}{n} B_n^{(k),\tau}(t) \big(Y_n^{(k),\tau}(t)\big)^\top \right\|_{\mathrm{F}} = \|B_n^{(k),\tau}(t)\|_n \|Y_n^{(k),\tau}(t)\|_n,$$

the preceding estimates give

$$\sup_n \left\| \|G_{n,L}^{(k)} - G_n^{(k)}\|_{L_t^2(\mathrm{F})} \right\|_{L^q} \le C_G(k,q) L^{-1/2}.$$

Define $\Pi_\tau(H) := H \min\{1, \tau/\|H\|_{L_t^2(\mathrm{F})}\}$, with multiplier one when $H = 0$. Since Π_τ is nonexpansive,

$$\left\| \Pi_\tau(G_{n,L}^{(k)}) - \Pi_\tau(G_n^{(k)}) \right\|_{L_t^2(\mathrm{F})} \le \left\| G_{n,L}^{(k)} - G_n^{(k)} \right\|_{L_t^2(\mathrm{F})}.$$

Since

$$V_{n,L}^{(k+1)} = V_{n,L}^{(k)} + \eta \Pi_\tau(G_{n,L}^{(k)}), \qquad V_n^{(k+1)} = V_n^{(k)} + \eta \Pi_\tau(G_n^{(k)}),$$

we obtain

$$\|V_{n,L}^{(k+1)} - V_n^{(k+1)}\|_{L_t^2(\mathrm{F})} \le \|V_{n,L}^{(k)} - V_n^{(k)}\|_{L_t^2(\mathrm{F})} + \eta \|G_{n,L}^{(k)} - G_n^{(k)}\|_{L_t^2(\mathrm{F})}.$$

The induction hypothesis and the bound on $\|G_{n,L}^{(k)} - G_n^{(k)}\|_{L_t^2(\mathrm{F})}$ therefore give (29) at $k+1$.

The induction therefore holds for every $k \le T$. Since the output estimate above holds for every $k \le T$ and T is fixed,

$$\max_{0 \le k \le T} \|f_{n,L}^{(k),\tau} - F_n^{(k),\tau}\|_{L^p} \le C_{\mathrm{D}}(T,p) L^{-1/2}.$$

This proves Proposition 4.3. □

C Continuous-Width Comparison

For notational convenience, write $Y_n^{(k)} := Y_n^{(k),\tau}$, $B_n^{(k)} := B_n^{(k),\tau}$, $F_n^{(k)} := F_n^{(k),\tau}$, and $\mathcal{J}_n^{(k)} := \mathcal{J}_n^{(k),\tau}$.

For $a, b \le T$, let

$$Q_{ab,n}(t) := \langle Y_n^{(a)}(t), Y_n^{(b)}(t) \rangle_n, \qquad R_{ab,n}(t) := \langle B_n^{(a)}(t), B_n^{(b)}(t) \rangle_n,$$

and

$$Q_{ab}(t) := \langle y^{(a)}(t), y^{(b)}(t) \rangle, \qquad R_{ab}(t) := \langle b^{(a)}(t), b^{(b)}(t) \rangle.$$

Write also $g_r := \gamma_r^\tau$.

We prove Proposition 4.4 in this section. We introduce surrogate states $\widetilde{Y}_n^{(k)}$ and $\widetilde{B}_n^{(k)}$ between the finite-width states $Y_n^{(k)}, B_n^{(k)}$ and the limiting states $y^{(k)}, b^{(k)}$. We first compare the surrogate quantities with their limiting values and then compare the finite-width states with the surrogate states. Backward estimates follow by depth reversal and transposition.

At the output level, the two comparisons meet through

$$F_n^{(k)} - f_\infty^{(k),\tau} = \langle \xi, Y_n^{(k)}(1) - \widetilde{Y}_n^{(k)}(1) \rangle_n + \left(\langle \xi, \widetilde{Y}_n^{(k)}(1) \rangle_n - f_\infty^{(k),\tau} \right),$$

where

$$Y_n^{(k)} \xrightarrow{n^{-1/2}} \widetilde{Y}_n^{(k)}, \qquad \langle \xi, \widetilde{Y}_n^{(k)}(1) \rangle_n \xrightarrow{n^{-1/2}} f_\infty^{(k),\tau}.$$

C.1 Surrogate construction

Define

$$\widetilde{Y}_n^{(k)}(t) = \Phi_{0,t} W_0 x - \sum_{r<k} g_r \int_0^t \Phi_{s,t} \widetilde{B}_n^{(r)}(s) Q_{rk}(s) \, \mathrm{d}s, \tag{32}$$

$$\widetilde{B}_n^{(k)}(t) = \Phi_{t,1}^\top \xi - \sum_{r<k} g_r \int_t^1 \Phi_{t,s}^\top \widetilde{Y}_n^{(r)}(s) R_{rk}(s) \, \mathrm{d}s. \tag{33}$$

In the Volterra system defining the surrogate states, the matrix-Brownian flow Φ, $W_0 x$, and ξ remain finite-width, while the coefficients in the integral terms are taken from the limiting model.

Proposition C.1 (Surrogate-to-limit comparison)**.** For every fixed T and every even $p \ge 2$, there exists $C_{\mathrm{sur}}(T,p) < \infty$, independent of n, such that

$$\max_{0 \le k \le T} \left\| \langle \xi, \widetilde{Y}_n^{(k)}(1) \rangle_n - f_\infty^{(k),\tau} \right\|_{L^p} \le C_{\mathrm{sur}}(T,p) n^{-1/2}. \tag{34}$$

In addition, for all $0 \le a, b \le T$,

$$\sup_{t\in[0,1]} \left\| \langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t) \rangle_n - Q_{ab}(t) \right\|_{L^p} \le C_{\mathrm{sur}}(T,p) n^{-1/2}, \tag{35}$$

$$\sup_{t\in[0,1]} \left\| \langle \widetilde{B}_n^{(a)}(t), \widetilde{B}_n^{(b)}(t) \rangle_n - R_{ab}(t) \right\|_{L^p} \le C_{\mathrm{sur}}(T,p) n^{-1/2}. \tag{36}$$

Proof. At $k = 0$,

$$\widetilde{Y}_n^{(0)}(t) = \Phi_{0,t} W_0 x, \qquad \widetilde{B}_n^{(0)}(t) = \Phi_{t,1}^\top \xi.$$

Now suppose that, for every $r < k$, there exist matrix-valued functions $K_{r,1}, K_{r,2}, L_{r,1}, L_{r,2}$ such that

$$\widetilde{Y}_n^{(r)}(t) = K_{r,1}(t) W_0 x + K_{r,2}(t)\xi, \qquad \widetilde{B}_n^{(r)}(t) = L_{r,1}(t) W_0 x + L_{r,2}(t)\xi. \tag{37}$$

Substituting the second identity in (37) into (32) gives

$$\widetilde{Y}_n^{(k)}(t) = \left[\Phi_{0,t} - \sum_{r<k} g_r \int_0^t \Phi_{s,t} L_{r,1}(s) Q_{rk}(s)\, \mathrm{d}s \right] W_0 x - \left[\sum_{r<k} g_r \int_0^t \Phi_{s,t} L_{r,2}(s) Q_{rk}(s)\, \mathrm{d}s \right] \xi.$$

Hence, $\widetilde{Y}_n^{(k)}$ has the same form as (37).

Likewise, substituting the first identity in (37) into (33) gives

$$\widetilde{B}_n^{(k)}(t) = -\left[\sum_{r<k} g_r \int_t^1 \Phi_{t,s}^\top K_{r,1}(s) R_{rk}(s)\, \mathrm{d}s \right] W_0 x + \left[\Phi_{t,1}^\top - \sum_{r<k} g_r \int_t^1 \Phi_{t,s}^\top K_{r,2}(s) R_{rk}(s)\, \mathrm{d}s \right] \xi.$$

Thus, $\widetilde{B}_n^{(k)}$ also has the form (37).

With (37), the surrogate output is

$$\langle \xi, \widetilde{Y}_n^{(k)}(1) \rangle_n = \frac{1}{n} \xi^\top K_{k,1}(1) W_0 x + \frac{1}{n} \xi^\top K_{k,2}(1) \xi.$$

Likewise,

$$\begin{aligned} \langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t) \rangle_n = {} & \frac{1}{n} (W_0 x)^\top K_{a,1}(t)^\top K_{b,1}(t) W_0 x + \frac{1}{n} (W_0 x)^\top K_{a,1}(t)^\top K_{b,2}(t) \xi \\ & + \frac{1}{n} \xi^\top K_{a,2}(t)^\top K_{b,1}(t) W_0 x + \frac{1}{n} \xi^\top K_{a,2}(t)^\top K_{b,2}(t) \xi. \end{aligned}$$

The corresponding formula for $\langle \widetilde{B}_n^{(a)}(t), \widetilde{B}_n^{(b)}(t) \rangle_n$ is obtained by replacing $K_{r,i}$ with $L_{r,i}$.

Conditioning on Φ, the matrices $K_{r,i}$ and $L_{r,i}$ are fixed, while $W_0 x$ and ξ remain independent centered Gaussian vectors with

$$\mathbb{E}[(W_0 x)(W_0 x)^\top] = q_0 I_n, \qquad \mathbb{E}[\xi \xi^\top] = I_n.$$

Therefore, for any such matrix K,

$$\mathbb{E}\left[\frac{1}{n} \xi^\top K \xi \,\middle|\, \Phi \right] = \frac{1}{n} \operatorname{tr}\big(K\, \mathbb{E}[\xi \xi^\top] \big) = \operatorname{tr}_n(K),$$

and similarly,

$$\mathbb{E}\left[\frac{1}{n} (W_0 x)^\top K (W_0 x) \,\middle|\, \Phi \right] = q_0 \operatorname{tr}_n(K).$$

The mixed terms vanish because $W_0 x$ and ξ are independent and centered:

$$\mathbb{E}\left[\frac{1}{n} (W_0 x)^\top K \xi \,\middle|\, \Phi \right] = 0.$$

Applying these identities to the expansions above gives

$$\mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big] = \mathrm{tr}_n\big(K_{k,2}(1)\big), \tag{38}$$

$$\mathbb{E}\Big[\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n \,\Big|\, \Phi\Big] = q_0\,\mathrm{tr}_n\big(K_{a,1}(t)^\top K_{b,1}(t)\big) + \mathrm{tr}_n\big(K_{a,2}(t)^\top K_{b,2}(t)\big), \tag{39}$$

$$\mathbb{E}\Big[\langle \widetilde{B}_n^{(a)}(t), \widetilde{B}_n^{(b)}(t)\rangle_n \,\Big|\, \Phi\Big] = q_0\,\mathrm{tr}_n\big(L_{a,1}(t)^\top L_{b,1}(t)\big) + \mathrm{tr}_n\big(L_{a,2}(t)^\top L_{b,2}(t)\big). \tag{40}$$

For the surrogate output, decompose

$$\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n - f_\infty^{(k),\tau} = \Big(\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n - \mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big]\Big) + \Big(\mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big] - f_\infty^{(k),\tau}\Big).$$

Using (37) at $t = 1$ and (38),

$$\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n - \mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big] = \underbrace{\frac{1}{n}\xi^\top K_{k,1}(1) W_0 x}_{\mathrm{I}} + \underbrace{\frac{1}{n}\xi^\top K_{k,2}(1)\xi - \mathrm{tr}_n\big(K_{k,2}(1)\big)}_{\mathrm{II}}.$$

The matrix-flow expansions established in the proof of Proposition E.2, together with (19), imply that, for every fixed T and $q \geq 2$,

$$\sup_{n\geq 1}\max_{\substack{0\leq k\leq T\\ i\in\{1,2\}}}\Big\|\sup_{t\in[0,1]}\|K_{k,i}(t)\|_{\mathrm{op}}\Big\|_{L^q} + \sup_{n\geq 1}\max_{\substack{0\leq k\leq T\\ i\in\{1,2\}}}\Big\|\sup_{t\in[0,1]}\|L_{k,i}(t)\|_{\mathrm{op}}\Big\|_{L^q} < \infty.$$

Indeed, for fixed T, only finitely many matrix-flow words and word lengths occur, and each product is controlled by (19) and Hölder's inequality.

Conditioning on Φ, Term I is a mixed Gaussian form in $W_0 x$ and ξ, while Term II is a centered Gaussian quadratic form in ξ. The operator-moment bounds above and the standard Gaussian moment estimates therefore give

$$\Big\|\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n - \mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big]\Big\|_{L^p} \lesssim n^{-1/2}.$$

Applying Proposition E.2 to (38) gives

$$\Big\|\mathbb{E}\Big[\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n \,\Big|\, \Phi\Big] - f_\infty^{(k),\tau}\Big\|_{L^p} \lesssim n^{-1/2}.$$

Together with the estimate for Terms I and II, this proves (34).

For the forward inner products, (37) and (39) give

$$\begin{aligned}
&\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n - \mathbb{E}\Big[\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n \,\Big|\, \Phi\Big] \\
&= \underbrace{\frac{1}{n}(W_0 x)^\top K_{a,1}(t)^\top K_{b,1}(t) W_0 x - q_0\,\mathrm{tr}_n\big(K_{a,1}(t)^\top K_{b,1}(t)\big)}_{\mathrm{I}} \\
&\quad + \underbrace{\frac{1}{n}(W_0 x)^\top K_{a,1}(t)^\top K_{b,2}(t)\xi}_{\mathrm{II}} + \underbrace{\frac{1}{n}\xi^\top K_{a,2}(t)^\top K_{b,1}(t) W_0 x}_{\mathrm{III}} \\
&\quad + \underbrace{\frac{1}{n}\xi^\top K_{a,2}(t)^\top K_{b,2}(t)\xi - \mathrm{tr}_n\big(K_{a,2}(t)^\top K_{b,2}(t)\big)}_{\mathrm{IV}}.
\end{aligned}$$

Terms I and IV are centered Gaussian quadratic forms, while Terms II and III are mixed Gaussian forms. The same Gaussian moment bounds used for the surrogate output therefore give

$$\sup_{t\in[0,1]}\Big\|\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n - \mathbb{E}\Big[\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n \,\Big|\, \Phi\Big]\Big\|_{L^p} \lesssim n^{-1/2}.$$

Applying Proposition E.2 to the two normalized traces in (39) gives

$$\sup_{t\in[0,1]}\Big\|\mathbb{E}\Big[\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t)\rangle_n \,\Big|\, \Phi\Big] - Q_{ab}(t)\Big\|_{L^p} \lesssim n^{-1/2}.$$

Combining the two estimates proves (35).

Replacing $K_{r,i}$ by $L_{r,i}$ in the decomposition above and using (40) gives, in the same way, (36). □

C.2 Approximation by the surrogate

The backward comparison follows from the same depth-reversal and transposition argument as in Appendix B. We therefore only prove the forward estimate below.

Proof of Proposition 4.4. We first prove the result for every even $p \geq 2$. We proceed by induction over the training step k. At $k = 0$, no gradient update happens, and

$$Y_n^{(0)}(t) = \widetilde{Y}_n^{(0)}(t) = \Phi_{0,t} W_0 x, \qquad B_n^{(0)}(t) = \widetilde{B}_n^{(0)}(t) = \Phi_{t,1}^\top \xi.$$

Proposition C.1 then gives the corresponding output and inner-product estimates.

Assume that, for every $r < k$ and every even $q \geq 2$,

$$\left\| \sup_{t\in[0,1]} \|Y_n^{(r)}(t) - \widetilde{Y}_n^{(r)}(t)\|_n \right\|_{L^q} + \left\| \sup_{t\in[0,1]} \|B_n^{(r)}(t) - \widetilde{B}_n^{(r)}(t)\|_n \right\|_{L^q} \lesssim n^{-1/2},$$

and

$$\|\gamma_n^{(r),\tau} - g_r\|_{L^q} \lesssim n^{-1/2}.$$

For the kth training step, define $V_n^{(k)}(s) := \sum_{r<k} \frac{\gamma_n^{(r),\tau}}{n} B_n^{(r)}(s)\big(Y_n^{(r)}(s)\big)^\top$. Then the finite-width forward state satisfies

$$Y_n^{(k)}(t) = \Phi_{0,t} W_0 x - \int_0^t \Phi_{s,t} V_n^{(k)}(s) Y_n^{(k)}(s)\, \mathrm{d}s.$$

Let $E_k(t) := Y_n^{(k)}(t) - \widetilde{Y}_n^{(k)}(t)$. Subtracting surrogate (32) from $Y_n^{(k)}(t)$ gives

$$E_k(t) = -\int_0^t \Phi_{s,t} V_n^{(k)}(s) E_k(s)\, \mathrm{d}s - \int_0^t \Phi_{s,t} \mathcal{F}_k(s)\, \mathrm{d}s, \tag{41}$$

where

$$\mathcal{F}_k(s) = \sum_{r<k} \Bigg[\underbrace{(\gamma_n^{(r),\tau} - g_r) B_n^{(r)}(s) \left\langle Y_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n}_{\mathrm{I}} + \underbrace{g_r \big(B_n^{(r)}(s) - \widetilde{B}_n^{(r)}(s)\big) \left\langle Y_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n}_{\mathrm{II}} + \underbrace{g_r \widetilde{B}_n^{(r)}(s) \left(\left\langle Y_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n - Q_{rk}(s) \right)}_{\mathrm{III}} \Bigg].$$

For Term III, write

$$\left\langle Y_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n - Q_{rk}(s) = \underbrace{\left\langle Y_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n}_{\mathrm{IIIa}} + \underbrace{\left(\left\langle \widetilde{Y}_n^{(r)}(s), \widetilde{Y}_n^{(k)}(s) \right\rangle_n - Q_{rk}(s) \right)}_{\mathrm{IIIb}}.$$

By Cauchy–Schwarz,

$$|\mathrm{IIIa}| \leq \left\| Y_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s) \right\|_n \left\| \widetilde{Y}_n^{(k)}(s) \right\|_n.$$

Since the induction has already established

$$\|\gamma_n^{(r),\tau} - g_r\|_{L^q} \lesssim n^{-1/2}, \qquad \left\| \sup_t \|B_n^{(r)}(t) - \widetilde{B}_n^{(r)}(t)\|_n \right\|_{L^q} \lesssim n^{-1/2},$$

and

$$\left\| \sup_t \|Y_n^{(r)}(t) - \widetilde{Y}_n^{(r)}(t)\|_n \right\|_{L^q} \lesssim n^{-1/2}, \qquad r < k,$$

Terms I, II, and IIIa are of order $n^{-1/2}$. Term IIIb is of the same order by Proposition C.1.

Proposition F.1, together with $\|V_n^{(r)}\|_{L_t^2(\mathrm{F})} \le r\eta\tau \le T\eta\tau$, gives uniform moments for $\|Y_n^{(r)}\|_n$ and $\|B_n^{(r)}\|_n$. Proposition C.1, applied with $a = b$, gives the corresponding bounds for $\|\widetilde{Y}_n^{(r)}\|_n$, $\|\widetilde{B}_n^{(r)}\|_n$, and $\|\widetilde{Y}_n^{(k)}\|_n$. Since g_r is deterministic, Hölder's inequality gives

$$\left\| \left(\int_0^1 \|\mathcal{F}_k(s)\|_n^2 \, \mathrm{d}s \right)^{1/2} \right\|_{L^{2p}} \lesssim n^{-1/2}.$$

Equation (41) has the form covered by the continuous stability estimate in Proposition F.2, with coefficient $V_n^{(k)}$ and forcing term $\mathcal{F}_k$. Since clipping gives $\|V_n^{(k)}\|_{L_t^2(\mathrm{F})} \le k\eta\tau$, the estimate above yields

$$\left\| \sup_{t\in[0,1]} \|Y_n^{(k)}(t) - \widetilde{Y}_n^{(k)}(t)\|_n \right\|_{L^p} \lesssim n^{-1/2}. \tag{42}$$

By the same depth-reversal and transposition argument,

$$\left\| \sup_{t\in[0,1]} \|B_n^{(k)}(t) - \widetilde{B}_n^{(k)}(t)\|_n \right\|_{L^p} \lesssim n^{-1/2}. \tag{43}$$

The output error decomposes as

$$F_n^{(k)} - f_\infty^{(k),\tau} = \underbrace{\left\langle \xi, Y_n^{(k)}(1) - \widetilde{Y}_n^{(k)}(1) \right\rangle_n}_{\text{(a)}} + \underbrace{\left(\left\langle \xi, \widetilde{Y}_n^{(k)}(1) \right\rangle_n - f_\infty^{(k),\tau} \right)}_{\text{(b)}}.$$

For Term (a), Cauchy–Schwarz, Hölder's inequality, and (42) give

$$\|(\mathrm{a})\|_{L^p} \le \|\|\xi\|_n\|_{L^{2p}} \left\| \sup_{t\in[0,1]} \|Y_n^{(k)}(t) - \widetilde{Y}_n^{(k)}(t)\|_n \right\|_{L^{2p}} \lesssim n^{-1/2}.$$

For Term (b), Proposition C.1 gives $\|(\mathrm{b})\|_{L^p} \lesssim n^{-1/2}$. Therefore,

$$\|F_n^{(k)} - f_\infty^{(k),\tau}\|_{L^p} \lesssim n^{-1/2}. \tag{44}$$

For $0 \le a, b \le k$,

$$\begin{aligned} Q_{ab,n}(t) - Q_{ab}(t) = &\left\langle Y_n^{(a)}(t) - \widetilde{Y}_n^{(a)}(t), Y_n^{(b)}(t) \right\rangle_n + \left\langle \widetilde{Y}_n^{(a)}(t), Y_n^{(b)}(t) - \widetilde{Y}_n^{(b)}(t) \right\rangle_n \\ &+ \left(\left\langle \widetilde{Y}_n^{(a)}(t), \widetilde{Y}_n^{(b)}(t) \right\rangle_n - Q_{ab}(t) \right). \end{aligned}$$

The first two terms are controlled by (42) and the uniform moments of $Y_n^{(a)}$ and $\widetilde{Y}_n^{(a)}$, while the last term is controlled by Proposition C.1. Hence

$$\sup_{t\in[0,1]} \|Q_{ab,n}(t) - Q_{ab}(t)\|_{L^p} \lesssim n^{-1/2}.$$

Using (43) in the same way gives

$$\sup_{t\in[0,1]} \|R_{ab,n}(t) - R_{ab}(t)\|_{L^p} \lesssim n^{-1/2}.$$

Since

$$\mathcal{J}_n^{(k)} = \int_0^1 Q_{kk,n}(t) R_{kk,n}(t) \, \mathrm{d}t, \qquad J_k^\tau = \int_0^1 Q_{kk}(t) R_{kk}(t) \, \mathrm{d}t,$$

we have

$$Q_{kk,n}R_{kk,n} - Q_{kk}R_{kk} = (Q_{kk,n} - Q_{kk})R_{kk,n} + Q_{kk}(R_{kk,n} - R_{kk}).$$

The bounds for $Q_{kk,n} - Q_{kk}$ and $R_{kk,n} - R_{kk}$ above, together with Hölder's inequality, therefore give

$$\|\mathcal{J}_n^{(k)} - J_k^\tau\|_{L^p} \lesssim n^{-1/2}.$$

We have now shown

$$\|F_n^{(k)} - f_\infty^{(k),\tau}\|_{L^p} + \|\mathcal{J}_n^{(k)} - J_k^\tau\|_{L^p} \lesssim n^{-1/2}.$$

Applying the preceding estimates at moment $2p$ and Lemma F.3 gives

$$\|\gamma_n^{(k),\tau} - g_k\|_{L^p} \lesssim n^{-1/2}. \tag{45}$$

This closes the induction at step k.

Since the induction holds for every $k \le T$, and T is fixed, (44) gives

$$\max_{0\le k\le T} \|F_n^{(k)} - f_\infty^{(k),\tau}\|_{L^p} \lesssim n^{-1/2}.$$

For an arbitrary finite $p \ge 2$, choose an even $p' \ge p$. The $L^{p'}$ estimate proved above implies the corresponding L^p estimate by monotonicity of L^p norms. This proves Proposition 4.4. □

D Commutativity in Probability

We prove Proposition 3.1 as a special case of the following finite-horizon result.

Proposition D.1 (Width and Depth Limits commute in probability during training)**.** Fix $T \in \mathbb{N}_0$. There exist deterministic limits $f_\infty^{(0)}, \dots, f_\infty^{(T)}$ such that

$$\max_{0\le k\le T} \left| f_{n,L}^{(k)} - f_\infty^{(k)} \right| \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{P}} 0.$$

Proof. Define the deterministic unclipped limiting trajectory using the recursion in Appendix E.1, with γ_k^τ replaced by $\gamma_k := \eta(f_\infty^{(k)} - y_\star)$. For fixed T, the same finite-step Volterra construction defines finite $f_\infty^{(k)}$ and J_k for $0 \le k \le T$.

For $0 \le k < T$, let $G_k := |f_\infty^{(k)} - y_\star|\sqrt{J_k}$, and choose $\tau_T := 1 + \max_{0\le k<T} G_k$.

By construction, clipping is inactive at every update of the deterministic τ_T-clipped trajectory. Hence, by induction over k,

$$f_\infty^{(k),\tau_T} = f_\infty^{(k)}, \qquad J_k^{\tau_T} = J_k, \qquad 0 \le k \le T. \tag{46}$$

For the finite network, define

$$G_{n,L}^{(k),\tau_T} := |f_{n,L}^{(k),\tau_T} - y_\star|\sqrt{J_{n,L}^{(k),\tau_T}}, \qquad 0 \le k < T.$$

By Theorem 4.1, the output satisfies

$$f_{n,L}^{(k),\tau_T} \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{L}_2} f_\infty^{(k),\tau_T}.$$

Moreover, the J-estimates established in the proofs of Propositions 4.3 and 4.4 give

$$J_{n,L}^{(k),\tau_T} \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{L}_2} J_k^{\tau_T}.$$

Together with (46),

$$G_{n,L}^{(k),\tau_T} \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{P}} G_k, \qquad 0 \le k < T.$$

Let

$$E_{n,L} := \left\{ G_{n,L}^{(k),\tau_T} < \tau_T \text{ for all } 0 \le k < T \right\}.$$

Since $\tau_T - G_k \ge 1$ for every $k < T$ and T is fixed, $\mathbb{P}(E_{n,L}) \longrightarrow 1$.

On $E_{n,L}$, the clipping factor is one at every step, so the updates are exactly the same as ordinary gradient descent. Since the clipped and ordinary trajectories start from the same initialization, induction over the training steps gives

$$f_{n,L}^{(k)} = f_{n,L}^{(k),\tau_T}, \qquad 0 \le k \le T, \quad \text{on } E_{n,L}.$$

Therefore, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{0\le k\le T} |f_{n,L}^{(k)} - f_{\infty}^{(k)}| > \varepsilon\right) \le \mathbb{P}(E_{n,L}^{c}) + \mathbb{P}\left(\max_{0\le k\le T} |f_{n,L}^{(k),\tau_T} - f_{\infty}^{(k),\tau_T}| > \varepsilon\right).$$

The first term tends to zero by the definition of $E_{n,L}$, while the second tends to zero by Theorem 4.1. This proves Proposition D.1. □

Lemma D.2 (Initialization limits)**.** As $\min\{n, L\} \to \infty$, $f_{n,L}^{(0)} \xrightarrow{\mathbb{L}_2} 0$, $J_{n,L}^{(0)} \xrightarrow{\mathbb{L}_2} q_0\mathrm{e}$. Moreover, for any fixed $\tau > |y_\star|\sqrt{q_0\mathrm{e}}$,

$$f_{\infty}^{(1),\tau} = \eta y_\star q_0 \mathrm{e}.$$

Proof. At initialization, clipping has not yet been applied, so we omit the superscript τ. Independence of ξ and $Y_L^{(0)}$ gives

$$\mathbb{E}|f_{n,L}^{(0)}|^2 = \frac{q_0(1 + L^{-1})^L}{n} \le \frac{q_0\mathrm{e}}{n}.$$

Hence $f_{n,L}^{(0)} \to 0$ in $\mathbb{L}_2$ as $\min\{n, L\} \to \infty$.

For the gradient term, the $k = 0$ estimates established in the proofs of Propositions 4.3 and 4.4 give $J_{n,L}^{(0)} - \mathcal{J}_n^{(0)} \xrightarrow[L\to\infty]{\mathbb{L}_2} 0$ uniformly in n, and $\mathcal{J}_n^{(0)} \xrightarrow[n\to\infty]{\mathbb{L}_2} J_0$, where J_0 denotes the limiting value at initialization. Therefore,

$$J_{n,L}^{(0)} \xrightarrow[\min\{n,L\}\to\infty]{\mathbb{L}_2} J_0.$$

It remains to identify J_0. The initialization covariance calculation of Hayou (2023) applies here with the identity activation. In their ReLU setting, the diagonal covariance equation contains the factor $\mathbb{E}[\mathrm{ReLU}(Z)^2] = 1/2$. Here $\phi(z) = z$, so $\mathbb{E}[\phi(Z)^2] = 1$, and the corresponding forward covariance satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} Q_{00}(t) = Q_{00}(t), \qquad Q_{00}(0) = q_0.$$

Thus, $Q_{00}(t) = q_0\mathrm{e}^t$. Applying the same calculation after depth reversal, with $R_{00}(1) = 1$, gives $R_{00}(t) = \mathrm{e}^{1-t}$. Consequently,

$$J_0 = \int_0^1 Q_{00}(t)R_{00}(t)\ \mathrm{d}t = q_0\mathrm{e}.$$

Now fix any $\tau > |y_\star|\sqrt{q_0\mathrm{e}}$. Since $f_{\infty}^{(0),\tau} = 0$ and $J_0 = q_0\mathrm{e}$,

$$|f_{\infty}^{(0),\tau} - y_\star|\sqrt{J_0} = |y_\star|\sqrt{q_0\mathrm{e}} < \tau.$$

Hence clipping is inactive at the first update of the limiting system, and $\gamma_0^\tau = -\eta y_\star$.

Applying (47) at $k = 1$ and taking the output inner product gives

$$f_{\infty}^{(1),\tau} = -\gamma_0^\tau \int_0^1 Q_{00}(t)R_{00}(t)\ \mathrm{d}t = -\gamma_0^\tau J_0 = \eta y_\star q_0 \mathrm{e}.$$

□

It remains to identify the one-step limit in Proposition 3.1. At initialization, $f_{\infty}^{(0)} = 0$ and $J_0 = q_0\mathrm{e}$, as established in Lemma D.2. Hence $\gamma_0 = -\eta y_\star$. Applying the limiting recursion at $k = 1$ gives

$$f_{\infty}^{(1)} = -\gamma_0 J_0 = \eta y_\star q_0 \mathrm{e}.$$

Taking $T = 1$ in Proposition D.1 therefore proves Proposition 3.1.

E Random-Matrix Estimates

The surrogate output in Appendix C has the form

$$\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n = \frac{1}{n}\xi^\top K_{k,1}(1)W_0 x + \frac{1}{n}\xi^\top K_{k,2}(1)\xi.$$

The estimates in this appendix establish the two comparisons

$$\frac{1}{n}\xi^\top K_{k,1}(1)W_0 x \xrightarrow{n^{-1/2}} 0, \qquad \frac{1}{n}\xi^\top K_{k,2}(1)\xi \xrightarrow{n^{-1/2}} \operatorname{tr}_n\big(K_{k,2}(1)\big) \xrightarrow{n^{-1}} f_\infty^{(k),\tau}.$$

We first establish an n^{-1} trace comparison for products of Φ and $\Phi^\top$, and then extend it to the matrix expressions $K_{k,i}$ and $L_{k,i}$ appearing in the surrogate system. This gives the $n^{-1/2}$ approximation of the surrogate output and the corresponding surrogate inner products required in Appendix C.

The trace comparison below connects the finite-width matrix flow with its infinite-width limit:

$$\begin{array}{ccccc} \text{matrix Brownian motion } M_t & \longrightarrow & \text{matrix flow } \Phi_{s,t} & \longrightarrow & \operatorname{tr}_n(\mathsf{w}_n) \\ \Downarrow & & \Downarrow & & \Downarrow \\ \text{circular Brownian motion } c_t & \longrightarrow & \text{circular flow } u_{s,t} & \longrightarrow & \varphi(\mathsf{w}_\infty). \end{array}$$

The upper row is the finite-width system introduced in Appendix B.1: M_t is the $n \times n$ matrix Brownian motion driving the continuous-depth dynamics, and $\Phi_{s,t}$ is its multiplicative flow. The lower row gives the corresponding infinite-width objects.

Formally, let $(\mathfrak{A}, \varphi)$ be a tracial noncommutative probability space carrying a standard circular Brownian motion c_t. Its multiplicative flow $u_{s,t}$ is defined by

$$u_{s,t} = 1 + \int_s^t \mathrm{d}c_r \, u_{s,r}, \qquad u_{s,s} = 1.$$

E.1 Deterministic limiting system

Let $(\mathfrak{A}, \varphi)$ be a tracial noncommutative probability space, and let $(c_t)_{t\in[0,1]}$ be a standard circular Brownian motion in $\mathfrak{A}$. For $0 \le s \le t \le 1$, let $u_{s,t}$ be the multiplicative circular flow defined by

$$u_{s,t} = 1 + \int_s^t \mathrm{d}c_r \, u_{s,r}, \qquad u_{s,s} = 1,$$

where the stochastic integral is the left free Itô integral with respect to the circular Brownian motion (c_t).

We regard $\mathbb{L}_2(\mathfrak{A}, \varphi)$ as a real Hilbert space with inner product $\langle a, b\rangle_{\mathcal{H}_0} := \operatorname{Re}\varphi(a^*b)$, and denote this real Hilbert space by $\mathcal{H}_0$. Set $\mathcal{H} := \mathcal{H}_0 \oplus \mathcal{H}_0$.

For $a \in \mathfrak{A}$, let L_a denote left multiplication by a on $\mathcal{H}_0$. Then $L_a^* = L_{a^*}$. Define $U_{s,t} := L_{u_{s,t}} \oplus L_{u_{s,t}}$ and set $P_{s,t}^{(0)} := U_{s,t}$.

Given the stages below k, define

$$V^{(k)}(t)h := \sum_{r<k} \gamma_r^\tau b^{(r)}(t)\langle y^{(r)}(t), h\rangle_{\mathcal{H}},$$

and let

$$P_{s,t}^{(k)} = U_{s,t} - \int_s^t U_{r,t}V^{(k)}(r)P_{s,r}^{(k)} \, \mathrm{d}r. \tag{47}$$

Set

$$y^{(k)}(t) := P_{0,t}^{(k)}(\sqrt{q_0}\,1, 0), \qquad b^{(k)}(t) := (P_{t,1}^{(k)})^*(0,1).$$

Then define

$$Q_{ab}(t) := \langle y^{(a)}(t), y^{(b)}(t)\rangle_{\mathcal{H}}, \qquad R_{ab}(t) := \langle b^{(a)}(t), b^{(b)}(t)\rangle_{\mathcal{H}},$$

$f_\infty^{(k),\tau} := \langle (0,1), y^{(k)}(1)\rangle_{\mathcal{H}}$ and $J_k^\tau := \int_0^1 Q_{kk}(t)R_{kk}(t)\,\mathrm{d}t$. For $k < T$, define γ_k^τ by the same clipping rule as in the finite-width model.

For every fixed k, (47) is a linear Volterra equation once the earlier stages are fixed, and hence has a unique solution.

E.2 MATRIX-BROWNIAN TRACE COMPARISON

We first establish the trace comparison underlying the width limit. At finite width, we consider normalized traces of products of the matrix flow $\Phi_{s,t}$ and its transpose. Their infinite-width limits are obtained from the corresponding products of the circular flow $u_{s,t}$ and its adjoint. At the trace level,

$$\underbrace{\mathrm{tr}_n\big(\Phi_{a_1,b_1}\,\Phi_{a_2,b_2}^{\top}\cdots\Phi_{a_d,b_d}\big)}_{\text{finite width}} \quad \xrightarrow{\;n^{-1}\;} \quad \underbrace{\varphi\big(u_{a_1,b_1}\,u_{a_2,b_2}^{*}\cdots u_{a_d,b_d}\big)}_{\text{infinite width}}.$$

The proposition below makes this comparison precise for arbitrary fixed products and uniformly over the depth intervals.

Proposition E.1 (Matrix–Brownian trace comparison)**.** Let w_n be a matrix-flow word [4]of length d and let w_∞ be its circular counterpart. For every even integer $p \geq 2$, there is a finite constant $C_{\mathrm{tr}}(d,p)$ such that

$$\|\mathrm{tr}_n(\mathsf{w}_n) - \varphi(\mathsf{w}_\infty)\|_{L^p} \leq C_{\mathrm{tr}}(d,p)n^{-1}. \tag{48}$$

The constant is independent of n and of all intervals appearing in the word.

Proof. Fix an auxiliary mesh $t_\ell := \ell/N$. For every interval $[a,b]$ appearing in the word, set $a_N := \lfloor Na\rfloor/N$ and $b_N := \lceil Nb\rceil/N$. The flow property gives $\Phi_{a_N,b_N} = \Phi_{b,b_N}\Phi_{a,b}\Phi_{a_N,a}$. Hence (19), (20), and (18), followed by a telescoping expansion and Hölder's inequality, imply

$$\sup_{n\geq 1}\|\mathrm{tr}_n(\mathsf{w}_n) - \mathrm{tr}_n(\mathsf{w}_{n,N})\|_{L^p} \leq C_{\mathrm{mat}}(d,p)N^{-1/2}, \tag{49}$$

where $\mathsf{w}_{n,N}$ is obtained by replacing every mesh-aligned factor Φ_{t_i,t_j} with

$$P_{i+1:j}^{n,N} := \big(I + N^{-1/2}G_{j,N}\big)\cdots\big(I + N^{-1/2}G_{i+1,N}\big),$$
$$G_{\ell,N} := \sqrt{N}(M_{t_\ell} - M_{t_{\ell-1}}).$$

The matrices $G_{\ell,N}$ are independent real Ginibre matrices with entries distributed as $\mathcal{N}(0,n^{-1})$.

We use the analogous Euler approximation on the circular side. Work in a von Neumann realization of $(\mathfrak{A},\varphi)$ and let $\bar{u}_{i,t}^N$ solve $\bar{u}_{i,t}^N = 1 + \int_{t_i}^t \mathrm{d}c_r\,\bar{u}_{i,\pi_N(r)}^N$, where $\pi_N(r)$ is the left endpoint of the mesh cell containing r. At the mesh points, $\bar{u}_{i,t_j}^N = (1+\Delta c_{j,N})\cdots(1+\Delta c_{i+1,N})$. The operator-norm Burkholder–Gundy inequality of Biane & Speicher (1998, Theorem 3.2.1), applied to the two semicircular components of c, gives uniform operator-norm bounds for u and $\bar{u}^N$. Applied to their difference, it gives

$$E_i(t) \leq \frac{C}{N} + C\int_{t_i}^t E_i(r)\,\mathrm{d}r, \qquad E_i(t) := \sup_{t_i\leq q\leq t}\|u_{t_i,q} - \bar{u}_{i,q}^N\|_\infty^2.$$

Gronwall's inequality and the same endpoint argument as above therefore give, for the corresponding circular Euler word $\mathsf{w}_{\infty,N}$,

$$|\varphi(\mathsf{w}_\infty) - \varphi(\mathsf{w}_{\infty,N})| \leq C_{\mathrm{free}}(d)N^{-1/2}. \tag{50}$$

It remains to compare the two Euler words uniformly in N. Put $g_{\ell,N} := \sqrt{N}(c_{t_\ell} - c_{t_{\ell-1}})$. The $g_{\ell,N}$ are freely independent standard circular elements, and the matrix and circular Euler factors are respectively $I + N^{-1/2}G_{\ell,N}$ and $1 + N^{-1/2}g_{\ell,N}$, possibly transposed or adjointed. Across $q \in \{1,p\}$ copies of a word, a fixed mesh index occurs at most qd times. Thus the total absolute Wick weight associated with that index is bounded by

$$A_{q,d}(N) := \sum_{m=0}^{\lfloor qd/2\rfloor}\binom{qd}{2m}(2m-1)!!\,N^{-m} \leq 1 + \frac{K_{q,d}}{N}, \qquad A_{q,d}(N)^N \leq \mathrm{e}^{K_{q,d}}. \tag{51}$$

[4]A matrix-flow word of length d is an ordered product of d factors $\Phi_{a,b}$ or $\Phi_{a,b}^{\top}$, where $0 \leq a \leq b \leq 1$; the intervals may overlap or coincide. Its circular counterpart is obtained by replacing $\Phi_{a,b}$ with $u_{a,b}$ and $\Phi_{a,b}^{\top}$ with $u_{a,b}^{*}$, without changing the order.

This is the required uniformity in N.

We now use the exact mixed-trace expansion for independent real Ginibre matrices and their transposes; see Redelmeier (2014, Lemmas 3.8 and 4.2). A pairing ρ of $2m$ Gaussian factors in one normalized trace contributes $n^{V(\rho)-m-1}$, where $V(\rho)$ is the number of free matrix indices. The associated closed surface has one face, m edges, and $V(\rho)$ vertices, so $V(\rho) - m + 1 \leq 2$ by Redelmeier (2014, Lemma 5.2). Equality holds precisely for the color-compatible noncrossing $*$-pairings identified in Redelmeier (2014, Lemma 5.10); these are exactly the pairings in the circular Wick formula. Every remaining diagram is therefore $O(n^{-1})$. Summing its absolute weight with (51) for $q = 1$ gives

$$|\mathbb{E}\,\mathrm{tr}_n(\mathsf{w}_{n,N}) - \varphi(\mathsf{w}_{\infty,N})| \leq C_{\mathrm{mean}}(d)n^{-1}. \tag{52}$$

Finally, set $Z_{n,N} := \mathrm{tr}_n(\mathsf{w}_{n,N})$. Because p is even, $\mathbb{E}|Z_{n,N} - \mathbb{E}Z_{n,N}|^p$ is the product of p centered trace copies. Inclusion–exclusion removes every pairing component containing only one trace. If a surviving component contains $r_j \geq 2$ traces, m_j pairs, and V_j free indices, then its n-exponent is $V_j - m_j - r_j$. The surface bound $V_j - m_j + r_j \leq 2$ makes this exponent at most $2 - 2r_j$. If there are c components, then $\sum_j r_j = p$ and $c \leq p/2$, so the total exponent is at most $2c - 2p \leq -p$. Using (51) with $q = p$ yields

$$\|Z_{n,N} - \mathbb{E}Z_{n,N}\|_{L^p} \leq C_{\mathrm{cen}}(d,p)n^{-1}. \tag{53}$$

Together with (52), this proves

$$\|\mathrm{tr}_n(\mathsf{w}_{n,N}) - \varphi(\mathsf{w}_{\infty,N})\|_{L^p} \leq C_{\mathrm{Eul}}(d,p)n^{-1},$$

uniformly in N. Combining this estimate with (49) and (50), and then letting $N \to \infty$, proves (48). □

E.3 Trace comparison for surrogate expressions

Appendix C writes the surrogate states as

$$\widetilde{Y}_n^{(k)}(t) = K_{k,1}(t)W_0x + K_{k,2}(t)\xi, \qquad \widetilde{B}_n^{(k)}(t) = L_{k,1}(t)W_0x + L_{k,2}(t)\xi.$$

Accordingly,

$$\langle \xi, \widetilde{Y}_n^{(k)}(1)\rangle_n = \frac{1}{n}\xi^\top K_{k,1}(1)W_0x + \frac{1}{n}\xi^\top K_{k,2}(1)\xi.$$

The conditional expectations in (38), (39), and (40) reduce the width comparison to trace estimates for the coefficient matrices. For the surrogate output, the required chain is

$$\frac{1}{n}\xi^\top K_{k,1}(1)W_0x \xrightarrow{n^{-1/2}} 0, \qquad \frac{1}{n}\xi^\top K_{k,2}(1)\xi \xrightarrow{n^{-1/2}} \mathrm{tr}_n\big(K_{k,2}(1)\big) \xrightarrow{n^{-1}} f_\infty^{(k),\tau}.$$

The proposition below gives the corresponding n^{-1} trace estimates for the matrices in (38)–(40).

Proposition E.2 (Trace comparison for surrogate coefficients)**.** Fix $T \geq 0$ and an even $p \geq 2$. There exists a finite constant $C_{\mathrm{st}}(T,p)$, independent of n, such that

$$\max_{0\leq k\leq T} \Big\| \mathrm{tr}_n\big(K_{k,2}(1)\big) - f_\infty^{(k),\tau} \Big\|_{L^p} \leq C_{\mathrm{st}}(T,p)n^{-1}. \tag{54}$$

Furthermore,

$$\max_{0\leq a,b\leq T} \sup_{t\in[0,1]} \big\| q_0\,\mathrm{tr}_n\big(K_{a,1}(t)^\top K_{b,1}(t)\big) + \mathrm{tr}_n\big(K_{a,2}(t)^\top K_{b,2}(t)\big) - Q_{ab}(t)\big\|_{L^p} \leq C_{\mathrm{st}}(T,p)n^{-1}, \tag{55}$$

and

$$\max_{0\leq a,b\leq T} \sup_{t\in[0,1]} \big\| q_0\,\mathrm{tr}_n\big(L_{a,1}(t)^\top L_{b,1}(t)\big) + \mathrm{tr}_n\big(L_{a,2}(t)^\top L_{b,2}(t)\big) - R_{ab}(t)\big\|_{L^p} \leq C_{\mathrm{st}}(T,p)n^{-1}. \tag{56}$$

Proof. We prove the forward estimates involving $K_{k,i}$. The corresponding bounds for $L_{k,i}$ follow from the depth-reversal and transposition argument of Section B.2.

We first establish the following structural claim: for every $k \le T$ and $i \in \{1,2\}$, $K_{k,i}(t)$ is a finite sum of deterministic multiple depth integrals of matrix-flow words.

Using the backward coefficient recursions in the forward recursions for $K_{k,i}$ gives

$$K_{k,1}(t) = \Phi_{0,t} + \sum_{q<r<k} g_r g_q \int_0^t \int_s^1 Q_{rk}(s) R_{qr}(v)\, \Phi_{s,t} \Phi_{s,v}^\top K_{q,1}(v)\, \mathrm{d}v\, \mathrm{d}s, \tag{57}$$

$$\begin{aligned} K_{k,2}(t) = &-\sum_{r<k} g_r \int_0^t Q_{rk}(s)\, \Phi_{s,t} \Phi_{s,1}^\top\, \mathrm{d}s \\ &+ \sum_{q<r<k} g_r g_q \int_0^t \int_s^1 Q_{rk}(s) R_{qr}(v)\, \Phi_{s,t} \Phi_{s,v}^\top K_{q,2}(v)\, \mathrm{d}v\, \mathrm{d}s. \end{aligned} \tag{58}$$

The claim follows by induction on k. At $k = 0$, $K_{0,1}(t) = \Phi_{0,t}$ and $K_{0,2}(t) = 0$. Assume that the claim holds for every $q < k$. In (57)–(58), each occurrence of $K_{q,i}$ is multiplied by $\Phi_{s,t}\Phi_{s,v}^\top$ and integrated over s and v. Hence every matrix-flow word in the expansion of $K_{q,i}$ is replaced by $\Phi_{s,t}\Phi_{s,v}^\top \mathsf{w}_n$, which is again a matrix-flow word. The remaining coefficients $g_r g_q Q_{rk}(s) R_{qr}(v)$ are deterministic. Thus, $K_{k,1}$ and $K_{k,2}$ are finite sums of deterministic multiple depth integrals of matrix-flow words.

Let $\bar{K}_{k,i}(t)$ denote the corresponding expression obtained by replacing every matrix-flow word with its circular word. For fixed T, only finitely many word lengths occur in these expansions. Since Proposition E.1 is uniform over the depth variables, Minkowski's inequality gives

$$\max_{\substack{0\le k\le T\\ i\in\{1,2\}}} \sup_{t\in[0,1]} \big\| \mathrm{tr}_n\big(K_{k,i}(t)\big) - \varphi\big(\bar{K}_{k,i}(t)\big) \big\|_{L^p} \le C_{\mathrm{st}}^{(1)}(T,p) n^{-1}. \tag{59}$$

The same argument applies to the products in (39). Expanding both coefficient matrices expresses $K_{a,i}(t)^\top K_{b,j}(t)$ as a finite sum of multiple depth integrals whose matrix-valued factors are again matrix-flow words. Hence

$$\max_{0\le a,b\le T} \sup_{t\in[0,1]} \big\| \mathrm{tr}_n\big(K_{a,i}(t)^\top K_{b,j}(t)\big) - \varphi\big(\bar{K}_{a,i}(t)^* \bar{K}_{b,j}(t)\big) \big\|_{L^p} \le C_{\mathrm{st}}^{(2)}(T,p) n^{-1}. \tag{60}$$

It remains to identify the circular coefficients with the limiting system. Define

$$\bar{y}^{(k)}(t) := \big(\sqrt{q_0}\, \bar{K}_{k,1}(t), \bar{K}_{k,2}(t)\big), \qquad \bar{b}^{(k)}(t) := \big(\sqrt{q_0}\, \bar{L}_{k,1}(t), \bar{L}_{k,2}(t)\big).$$

Replacing Φ by u and $\Phi^\top$ by u^* in the coefficient recursions shows that $\bar{y}^{(k)}$ and $\bar{b}^{(k)}$ satisfy (47). By uniqueness, $\bar{y}^{(k)} = y^{(k)}$ and $\bar{b}^{(k)} = b^{(k)}$.

Consequently,

$$f_\infty^{(k),\tau} = \mathrm{Re}\, \varphi\big(\bar{K}_{k,2}(1)\big),$$

and

$$Q_{ab}(t) = q_0\, \mathrm{Re}\, \varphi\big(\bar{K}_{a,1}(t)^* \bar{K}_{b,1}(t)\big) + \mathrm{Re}\, \varphi\big(\bar{K}_{a,2}(t)^* \bar{K}_{b,2}(t)\big).$$

Similarly,

$$R_{ab}(t) = q_0\, \mathrm{Re}\, \varphi\big(\bar{L}_{a,1}(t)^* \bar{L}_{b,1}(t)\big) + \mathrm{Re}\, \varphi\big(\bar{L}_{a,2}(t)^* \bar{L}_{b,2}(t)\big).$$

Therefore (59) and (60) give (54), (55), and (56). □

F Stability under Clipped Training

The training update at depth t depends on both the forward and backward states. At initialization,

$$Y^{(0)}(t) = \Phi_{0,t} Y_0, \qquad B^{(0)}(t) = \Phi_{t,1}^\top \xi.$$

Thus $Y^{(0)}(t)$ depends on the matrix-Brownian increments over $[0,t]$, whereas $B^{(0)}(t)$ depends on those over $[t,1]$. Let $\mathcal{F}_t^M := \sigma(M_s : 0 \le s \le t)$.

Since

$$V^{(k)}(t) = \sum_{r<k} \frac{\gamma_r^\tau}{n} B^{(r)}(t)\big(Y^{(r)}(t)\big)^\top$$

contains the backward states, it is not $\mathcal{F}_t^M$-measurable. Consequently, the predictability condition in Lemma B.1 is not available during training.

We therefore use the Volterra identity

$$\Phi^V_{s,t} = \Phi_{s,t} - \int_s^t \Phi_{r,t} V(r) \Phi^V_{s,r} \, \mathrm{d}r. \tag{61}$$

Taking operator norms gives

$$\|\Phi^V_{s,t}\|_{\mathrm{op}} \le \|\Phi_{s,t}\|_{\mathrm{op}} + \int_s^t \|\Phi_{r,t}\|_{\mathrm{op}} \|V(r)\|_{\mathrm{op}} \|\Phi^V_{s,r}\|_{\mathrm{op}} \, \mathrm{d}r,$$

where the factors involving Φ are controlled by (19). Thus it remains to control the size of the cumulative training update. By global clipping, $\|V^{(k)}\|_{L^2_t(\mathrm{F})} \le k\eta\tau$. Section F.1 uses this bound to control $\Phi^{V^{(k)}}$ and hence the moments of $Y^{(k)}$ and $B^{(k)}$. Section F.2 shows that the $O(L^{-1/2})$ and $O(n^{-1/2})$ comparison rates are preserved when forming the next clipped update.

At stage k, the induction closes through

$$V^{(k)} \longrightarrow (Y^{(k)}, B^{(k)}, f^{(k)}, J^{(k)}) \longrightarrow \gamma^{(k)} \longrightarrow V^{(k+1)}.$$

F.1 Moment control during training

Proposition F.1 (Uniform bounds for trained flows)**.** Let $\Lambda < \infty$ and suppose that $\|V\|_{L^2_t(\mathrm{F})} \le \Lambda$ a.s. Let $\Phi^V_{s,t}$ be the trained propagator defined by

$$\Phi^V_{s,t} = \Phi_{s,t} - \int_s^t \Phi_{r,t} V(r) \Phi^V_{s,r} \, \mathrm{d}r.$$

Then this equation has a unique solution. Moreover, for every $p \ge 2$, there exists $C_{\mathrm{train}}(p,\Lambda) < \infty$, independent of n, such that

$$\sup_{0\le s\le 1} \Big\| \sup_{s\le t\le 1} \|\Phi^V_{s,t}\|_{\mathrm{op}} \Big\|_{L^p} \le C_{\mathrm{train}}(p,\Lambda). \tag{62}$$

The same bound holds for the reversed and transposed flow.
For the discrete trained states, for every fixed T and $p \ge 2$, there exists $C_{\mathrm{state}}(T,p) < \infty$, independent of n and L, such that

$$\max_{0\le k\le T} \Big\| \max_{0\le j\le L} \|Y^{(k),\tau}_{n,L}(t_j)\|_n \Big\|_{L^p} + \max_{0\le k\le T} \Big\| \max_{0\le j\le L} \|B^{(k),\tau}_{n,L}(t_j)\|_n \Big\|_{L^p} \le C_{\mathrm{state}}(T,p).$$

Proof. We first prove pathwise well-posedness. Since

$$\|V(r)\|_{\mathrm{op}} \le \|V(r)\|_{\mathrm{F}},$$

Cauchy–Schwarz gives

$$\int_0^1 \|V(r)\|_{\mathrm{op}} \, \mathrm{d}r \le \|V\|_{L^2_t(\mathrm{F})} \le \Lambda. \tag{63}$$

Fix a realization for which (63) holds. Since $(s,t) \mapsto \Phi_{s,t}$ is continuous on $\{0 \le s \le t \le 1\}$,

$$K_\Phi := \sup_{0\le s\le t\le 1} \|\Phi_{s,t}\|_{\mathrm{op}} < \infty.$$

Iterating the Volterra equation gives

$$\Phi^V_{s,t} = \Phi_{s,t} + \sum_{m\geq 1} (-1)^m I_m(s,t),$$

where

$$I_m(s,t) := \int_{s<r_1<\cdots<r_m<t} \Phi_{r_m,t} V(r_m) \Phi_{r_{m-1},r_m} \cdots V(r_1) \Phi_{s,r_1} \, \mathrm{d}\mathbf{r}.$$

By submultiplicativity of the operator norm,

$$\|I_m(s,t)\|_{\mathrm{op}} \leq K_\Phi^{m+1} \int_{s<r_1<\cdots<r_m<t} \prod_{i=1}^m \|V(r_i)\|_{\mathrm{op}} \, \mathrm{d}\mathbf{r}.$$

The integrand is symmetric in $r_1, \ldots, r_m$, so

$$\|I_m(s,t)\|_{\mathrm{op}} \leq \frac{K_\Phi^{m+1}}{m!} \left(\int_s^t \|V(r)\|_{\mathrm{op}} \, \mathrm{d}r \right)^m \leq K_\Phi \frac{(K_\Phi \Lambda)^m}{m!}.$$

Hence, the Volterra series converges absolutely in operator norm.

For uniqueness, let $\Psi_{s,t}$ and $\widetilde{\Psi}_{s,t}$ be two solutions and set $D_{s,t} := \Psi_{s,t} - \widetilde{\Psi}_{s,t}$. Then

$$D_{s,t} = -\int_s^t \Phi_{r,t} V(r) D_{s,r} \, \mathrm{d}r.$$

Therefore,

$$\|D_{s,t}\|_{\mathrm{op}} \leq K_\Phi \int_s^t \|V(r)\|_{\mathrm{op}} \, \|D_{s,r}\|_{\mathrm{op}} \, \mathrm{d}r.$$

Gronwall's inequality gives $D_{s,t} = 0$ for all $s \leq t$. Thus, the Volterra equation has a unique pathwise solution.

For $m \geq 1$, submultiplicativity of the operator norm gives

$$\sup_{t\geq s} \|I_m(s,t)\|_{\mathrm{op}} \leq \int_{s<r_1<\cdots<r_m<1} \sup_{t\geq r_m} \|\Phi_{r_m,t}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,r_{i+1}}\|_{\mathrm{op}} \|\Phi_{s,r_1}\|_{\mathrm{op}} \prod_{i=1}^m \|V(r_i)\|_{\mathrm{op}} \, \mathrm{d}\mathbf{r}.$$

By Cauchy–Schwarz in $r_1, \ldots, r_m$,

$$\sup_{t\geq s} \|I_m(s,t)\|_{\mathrm{op}} \leq \left(\int_{s<r_1<\cdots<r_m<1} \left[\sup_{u\geq r_m} \|\Phi_{r_m,u}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,r_{i+1}}\|_{\mathrm{op}} \|\Phi_{s,r_1}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2} \times \left(\int_{s<r_1<\cdots<r_m<1} \prod_{i=1}^m \|V(r_i)\|_{\mathrm{op}}^2 \, \mathrm{d}\mathbf{r} \right)^{1/2}.$$

Since $\|V(r)\|_{\mathrm{op}} \leq \|V(r)\|_{\mathrm{F}}$ and the integrand is symmetric in $r_1, \ldots, r_m$,

$$\int_{s<r_1<\cdots<r_m<1} \prod_{i=1}^m \|V(r_i)\|_{\mathrm{op}}^2 \, \mathrm{d}\mathbf{r} \leq \frac{1}{m!} \left(\int_s^1 \|V(r)\|_{\mathrm{F}}^2 \, \mathrm{d}r \right)^m \leq \frac{\Lambda^{2m}}{m!}.$$

Hence

$$\sup_{t\geq s} \|I_m(s,t)\|_{\mathrm{op}} \leq \frac{\Lambda^m}{\sqrt{m!}} \left(\int_{s<r_1<\cdots<r_m<1} \left[\sup_{u\geq r_m} \|\Phi_{r_m,u}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,r_{i+1}}\|_{\mathrm{op}} \|\Phi_{s,r_1}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2}.$$

For fixed $s < r_1 < \cdots < r_m$, the factors

$$\Phi_{s,r_1}, \, \Phi_{r_1,r_2}, \ldots, \, \Phi_{r_{m-1},r_m}, \qquad \sup_{u\geq r_m} \|\Phi_{r_m,u}\|_{\mathrm{op}},$$

depend on disjoint matrix-Brownian increments. Hence they are independent, and (19) gives

$$\left\| \sup_{u \ge r_m} \|\Phi_{r_m,u}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,r_{i+1}}\|_{\mathrm{op}} \|\Phi_{s,r_1}\|_{\mathrm{op}} \right\|_{L^p} \le C_{\mathrm{base}}(p)^{m+1}.$$

Since $p \ge 2$, Minkowski's inequality yields

$$\left\| \left(\int_{s<r_1<\cdots<r_m<1} \left[\sup_{u \ge r_m} \|\Phi_{r_m,u}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,r_{i+1}}\|_{\mathrm{op}} \|\Phi_{s,r_1}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2} \right\|_{L^p} \le \frac{C_{\mathrm{base}}(p)^{m+1}}{\sqrt{m!}}.$$

Combining this with the preceding estimate gives

$$\left\| \sup_{t \ge s} \|I_m(s,t)\|_{\mathrm{op}} \right\|_{L^p} \le C_{\mathrm{base}}(p) \frac{\big(C_{\mathrm{base}}(p)\Lambda\big)^m}{m!}. \tag{64}$$

Using the Volterra expansion and the triangle inequality in L^p,

$$\begin{aligned} \left\| \sup_{t \ge s} \|\Phi^V_{s,t}\|_{\mathrm{op}} \right\|_{L^p} &\le \left\| \sup_{t \ge s} \|\Phi_{s,t}\|_{\mathrm{op}} \right\|_{L^p} + \sum_{m \ge 1} \left\| \sup_{t \ge s} \|I_m(s,t)\|_{\mathrm{op}} \right\|_{L^p} \\ &\le C_{\mathrm{base}}(p) \sum_{m \ge 0} \frac{\big(C_{\mathrm{base}}(p)\Lambda\big)^m}{m!} \\ &= C_{\mathrm{base}}(p) \exp\big(C_{\mathrm{base}}(p)\Lambda\big). \end{aligned}$$

The right-hand side is independent of s and n, which proves (62).

For the backward flow, set $u := 1 - t$ and define

$$\check{M}_u := (M_1 - M_{1-u})^\top, \qquad \check{V}(u) := V(1-u)^\top.$$

Then $\check{M}$ has the same law as M, and $\|\check{V}\|_{L^2_t(\mathrm{F})} = \|V\|_{L^2_t(\mathrm{F})} \le \Lambda$. Therefore, we can apply the same estimate again to the reversed and transposed flow. Applying these bounds to Y_0 and ξ, together with their Gaussian moment bounds, gives the corresponding uniform moments for the forward and backward states.

The same argument gives the discrete state bounds used in Section B.2. By (27), for every $k \le T$,

$$\frac{1}{L} \sum_{\ell=1}^{L} \|V^{(k)}_{n,L}(t_\ell)\|_{\mathrm{F}}^2 \le (T\eta\tau)^2.$$

From

$$Y^{(k),\tau}_{n,L}(t_j) = S^{n,L}_{1:j} Y_0 - \frac{1}{L} \sum_{\ell_2=1}^{j} S^{n,L}_{\ell_2+1:j} V^{(k)}_{n,L}(t_{\ell_2}) Y^{(k),\tau}_{n,L}(t_{\ell_2-1}),$$

substituting the same identity for $Y^{(k),\tau}_{n,L}(t_{\ell_2-1})$ gives

$$\begin{aligned} Y^{(k),\tau}_{n,L}(t_j) = S^{n,L}_{1:j} Y_0 &- \frac{1}{L} \sum_{\ell_2=1}^{j} S^{n,L}_{\ell_2+1:j} V^{(k)}_{n,L}(t_{\ell_2}) S^{n,L}_{1:\ell_2-1} Y_0 \\ &+ \frac{1}{L^2} \sum_{\ell_1<\ell_2\le j} S^{n,L}_{\ell_2+1:j} V^{(k)}_{n,L}(t_{\ell_2}) S^{n,L}_{\ell_1+1:\ell_2-1} V^{(k)}_{n,L}(t_{\ell_1}) Y^{(k),\tau}_{n,L}(t_{\ell_1-1}). \end{aligned}$$

Continuing this substitution gives the finite expansion

$$Y^{(k),\tau}_{n,L}(t_j) = \sum_{m=0}^{j} (-1)^m Y^{(k)}_{j,m},$$

where $Y_{j,0}^{(k)} = S_{1:j}^{n,L} Y_0$, and, for $m \geq 1$,

$$Y_{j,m}^{(k)} = \frac{1}{L^m} \sum_{1 \leq \ell_1 < \cdots < \ell_m \leq j} S_{\ell_m+1:j}^{n,L} V_{n,L}^{(k)}(t_{\ell_m}) S_{\ell_{m-1}+1:\ell_m-1}^{n,L} \cdots V_{n,L}^{(k)}(t_{\ell_1}) S_{1:\ell_1-1}^{n,L} Y_0.$$

Thus the order-m term contains m update factors, with base products over the layer blocks between successive update indices.

By Cauchy–Schwarz over $1 \leq \ell_1 < \cdots < \ell_m \leq L$ and (27),

$$\frac{1}{L^m} \sum_{\ell_1 < \cdots < \ell_m} \prod_{i=1}^{m} \|V_{n,L}^{(k)}(t_{\ell_i})\|_{\mathrm{F}}^2 \leq \frac{(T\eta\tau)^{2m}}{m!}.$$

For each ordered tuple, the base products depend on disjoint blocks of residual matrices and are therefore independent. Using (19) and the same Cauchy–Schwarz and Minkowski argument as above gives

$$\left\| \max_{0 \leq j \leq L} \|Y_{j,m}^{(k)}\|_n \right\|_{L^p} \leq C_{\mathrm{base}}(2p) \frac{\big(C_{\mathrm{base}}(2p) T\eta\tau\big)^m}{m!} \, \| \|Y_0\|_n \|_{L^{2p}} .$$

Summing over m yields

$$\left\| \max_{0 \leq j \leq L} \|Y_{n,L}^{(k),\tau}(t_j)\|_n \right\|_{L^p} \leq C_{\mathrm{base}}(2p) \exp\big(C_{\mathrm{base}}(2p) T\eta\tau\big) \, \| \|Y_0\|_n \|_{L^{2p}} .$$

The Gaussian initialization gives $\sup_n \| \|Y_0\|_n \|_{L^{2p}} < \infty$. Applying the same argument after reversal and transposition, with terminal state ξ, gives the corresponding bound for $B_{n,L}^{(k),\tau}$. Taking the maximum over $k \leq T$ proves the discrete state bound. □

F.2 Error propagation during training

Let H_n denote $\mathbb{R}^n$ equipped with the normalized norm $\|z\|_n := n^{-1/2}\|z\|_2$.

Proposition F.2 (Bounds for the discrete and continuous error equations)**.** Fix $p \geq 2$ and $\Lambda < \infty$, and let V satisfy $\|V\|_{L_t^2(\mathrm{F})} \leq \Lambda$ a.s.
For integers $1 \leq a \leq b \leq L$, consider the discrete equation

$$H_{a,b} = \mathcal{R}_{a,b} - \sum_{j=a}^{b} \int_{I_j} \Phi_{r,t_b} V(r) H_{a,j-1} \, \mathrm{d}r, \qquad H_{a,a-1} = R_{a,a-1}, \tag{65}$$

there exists $C_{\mathrm{stab}}(p, \Lambda) < \infty$, independent of n, L, and a, such that

$$\left\| \max_{a \leq b \leq L} \|H_{a,b}\|_n . \right\|_{L^p} \leq C_{\mathrm{stab}}(p, \Lambda) \left\| \max_{b \geq a-1} \|\mathcal{R}_{a,b}\|_n \right\|_{L^{2p}}. \tag{66}$$

The same estimate holds after reversal and transposition.
For the continuous equation

$$E(t) = -\int_0^t \Phi_{s,t} V(s) E(s) \, \mathrm{d}s - \int_0^t \Phi_{s,t} F(s) \, \mathrm{d}s, \tag{67}$$

we have

$$\left\| \sup_{t \in [0,1]} \|E(t)\|_n \right\|_{L^p} \leq C_{\mathrm{stab}}(p, \Lambda) \|F\|_{L^{2p}(\Omega; L_t^2(H_n))}. \tag{68}$$

The same estimate holds after reversal and transposition.

Proof. We first consider the continuous equation. Define $\widetilde{E}(t) := -\int_0^t \Phi_{s,t}^V F(s) \, \mathrm{d}s$. Using the Volterra identity (61),

$$\widetilde{E}(t) = -\int_0^t \Phi_{s,t} F(s) \, \mathrm{d}s + \int_0^t \int_s^t \Phi_{r,t} V(r) \Phi_{s,r}^V F(s) \, \mathrm{d}r \, \mathrm{d}s.$$

The double integral is over $0 \le s \le r \le t$. By Fubini's theorem,

$$\int_0^t \int_s^t \Phi_{r,t} V(r) \Phi^V_{s,r} F(s)\, \mathrm{d}r\, \mathrm{d}s = \int_0^t \Phi_{r,t} V(r) \left(\int_0^r \Phi^V_{s,r} F(s)\, \mathrm{d}s \right) \mathrm{d}r.$$

By the definition of $\widetilde{E}$,

$$\int_0^r \Phi^V_{s,r} F(s)\, \mathrm{d}s = -\widetilde{E}(r).$$

Hence

$$\widetilde{E}(t) = -\int_0^t \Phi_{s,t} F(s)\, \mathrm{d}s - \int_0^t \Phi_{r,t} V(r) \widetilde{E}(r)\, \mathrm{d}r.$$

Thus $\widetilde{E}$ satisfies (67). The same Gronwall argument used above gives uniqueness, and therefore

$$E(t) = -\int_0^t \Phi^V_{s,t} F(s)\, \mathrm{d}s. \tag{69}$$

It follows that

$$\sup_{t\in[0,1]} \|E(t)\|_n \le \left(\int_0^1 \sup_{t\ge s} \|\Phi^V_{s,t}\|^2_{\mathrm{op}}\, \mathrm{d}s \right)^{1/2} \|F\|_{L^2_t(H_n)}.$$

By Hölder's inequality,

$$\left\| \sup_{t\in[0,1]} \|E(t)\|_n \right\|_{L^p} \le \left\| \left(\int_0^1 \sup_{t\ge s} \|\Phi^V_{s,t}\|^2_{\mathrm{op}}\, \mathrm{d}s \right)^{1/2} \right\|_{L^{2p}} \times \|F\|_{L^{2p}(\Omega; L^2_t(H_n))}.$$

Minkowski's inequality and Proposition F.1 give

$$\left\| \left(\int_0^1 \sup_{t\ge s} \|\Phi^V_{s,t}\|^2_{\mathrm{op}}\, \mathrm{d}s \right)^{1/2} \right\|_{L^{2p}} \le C_{\mathrm{train}}(2p, \Lambda).$$

Hence (68) holds with $C_{\mathrm{cont}}(p, \Lambda) := C_{\mathrm{train}}(2p, \Lambda)$.

We next consider the discrete equation. Picard iteration of (65) gives a finite expansion whose order-m terms are indexed by $a \le j_1 < \cdots < j_m \le b$. Let $H^{[m]}_{a,b}$ denote the term containing m factors of V. By submultiplicativity,

$$\max_{b\ge a} \|H^{[m]}_{a,b}\|_n \le \max_{c\ge a-1} \|R_{a,c}\|_n \sum_{a\le j_1<\cdots<j_m\le L} \int_{I_{j_1}\times\cdots\times I_{j_m}} \sup_{b\ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}} \times \prod_{i=1}^{m-1} \|\Phi_{r_i,t_{j_{i+1}-1}}\|_{\mathrm{op}} \prod_{i=1}^{m} \|V(r_i)\|_{\mathrm{op}}\, \mathrm{d}\mathbf{r}.$$

Applying Cauchy–Schwarz gives

$$\max_{b\ge a} \|H^{[m]}_{a,b}\|_n \le \left(\sum_{a\le j_1<\cdots<j_m\le L} \int_{I_{j_1}\times\cdots\times I_{j_m}} \left[\sup_{b\ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,t_{j_{i+1}-1}}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2} \times \left(\sum_{a\le j_1<\cdots<j_m\le L} \int_{I_{j_1}\times\cdots\times I_{j_m}} \prod_{i=1}^{m} \|V(r_i)\|^2_{\mathrm{op}}\, \mathrm{d}\mathbf{r} \right)^{1/2} \max_{c\ge a-1} \|R_{a,c}\|_n.$$

Since $r_i \in I_{j_i}$ and $j_1 < \cdots < j_m$, we have $r_1 < \cdots < r_m$. The sets $I_{j_1} \times \cdots \times I_{j_m}$ corresponding to distinct index tuples are disjoint. Therefore,

$$\sum_{a\le j_1<\cdots<j_m\le L} \int_{I_{j_1}\times\cdots\times I_{j_m}} \prod_{i=1}^{m} \|V(r_i)\|^2_{\mathrm{op}}\, \mathrm{d}\mathbf{r} \le \int_{0<r_1<\cdots<r_m<1} \prod_{i=1}^{m} \|V(r_i)\|^2_{\mathrm{op}}\, \mathrm{d}\mathbf{r}.$$

Using $\|V(r)\|_{\mathrm{op}} \le \|V(r)\|_{\mathrm{F}}$ and symmetry in $r_1, \dots, r_m$,

$$\int_{0<r_1<\cdots<r_m<1} \prod_{i=1}^{m} \|V(r_i)\|_{\mathrm{op}}^2 \, \mathrm{d}\mathbf{r} \le \frac{1}{m!} \left(\int_0^1 \|V(r)\|_{\mathrm{F}}^2 \, \mathrm{d}r \right)^m \le \frac{\Lambda^{2m}}{m!}.$$

For fixed $j_1 < \cdots < j_m$ and $r_i \in I_{j_i}$, the factors $\Phi_{r_i,t_{j_{i+1}-1}}$, $1 \le i < m$, depend on Brownian increments before r_{i+1}, while $\sup_{b \ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}}$ depends only on increments after r_m. These intervals are disjoint, so the corresponding random variables are independent. By (19),

$$\left\| \sup_{b \ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,t_{j_{i+1}-1}}\|_{\mathrm{op}} \right\|_{L^{2p}} \le C_{\mathrm{base}}(2p)^m.$$

Since each I_j has length L^{-1},

$$\sum_{a \le j_1 < \cdots < j_m \le L} |I_{j_1}| \cdots |I_{j_m}| = \binom{L-a+1}{m} L^{-m} \le \frac{1}{m!}.$$

Minkowski's inequality therefore gives

$$\left\| \left(\sum_{a \le j_1 < \cdots < j_m \le L} \int_{I_{j_1} \times \cdots \times I_{j_m}} \left[\sup_{b \ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,t_{j_{i+1}-1}}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2} \right\|_{L^{2p}} \le \frac{C_{\mathrm{base}}(2p)^m}{\sqrt{m!}}.$$

Combining the preceding estimates,

$$\max_{b \ge a} \|H_{a,b}^{[m]}\|_n \le \frac{\Lambda^m}{\sqrt{m!}} \left(\sum_{a \le j_1 < \cdots < j_m \le L} \int_{I_{j_1} \times \cdots \times I_{j_m}} \left[\sup_{b \ge j_m} \|\Phi_{r_m,t_b}\|_{\mathrm{op}} \prod_{i=1}^{m-1} \|\Phi_{r_i,t_{j_{i+1}-1}}\|_{\mathrm{op}} \right]^2 \mathrm{d}\mathbf{r} \right)^{1/2} \times \max_{c \ge a-1} \|R_{a,c}\|_n.$$

Taking L^p norms and applying Hölder's inequality,

$$\left\| \max_{b \ge a} \|H_{a,b}^{[m]}\|_n \right\|_{L^p} \le \frac{\big(C_{\mathrm{base}}(2p)\Lambda\big)^m}{m!} \left\| \max_{c \ge a-1} \|R_{a,c}\|_n \right\|_{L^{2p}}.$$

Since $H_{a,b}^{[0]} = \mathcal{R}_{a,b}$, summing over m gives

$$\left\| \max_{a \le b \le L} \|H_{a,b}\|_n . \right\|_{L^p} \le \exp\big(C_{\mathrm{base}}(2p)\Lambda\big) \left\| \max_{c \ge a-1} \|R_{a,c}\|_n \right\|_{L^{2p}}.$$

Thus (66) holds with $C_{\mathrm{disc}}(p,\Lambda) := \exp(C_{\mathrm{base}}(2p)\Lambda)$.

Reversal and transposition preserve the Brownian law, the $L_t^2(\mathrm{F})$ bound on V, and the relevant norms. Hence the same estimates hold for the backward equations. Finally, taking $C_{\mathrm{stab}}(p,\Lambda) := \max\{C_{\mathrm{cont}}(p,\Lambda), C_{\mathrm{disc}}(p,\Lambda)\}$ proves the proposition. □

Lemma F.3 (Clipped coefficient comparison)**.** Fix $k \le T$ and $p \ge 2$. Suppose

$$\|F_n^{(k)} - f_\infty^{(k),\tau}\|_{L^{2p}} + \|\mathcal{J}_n^{(k)} - J_k^\tau\|_{L^{2p}} \le C_k n^{-1/2},$$

and $\sup_n \|F_n^{(k)}\|_{L^{2p}} < \infty$. Then

$$\|\gamma_n^{(k),\tau} - g_k\|_{L^p} \le C_{k,p} n^{-1/2}.$$

Proof. We first establish a deterministic local bound for the clipping map.

If $|f_\infty^{(k),\tau} - y_\star|\sqrt{J_k^\tau} < \tau$, continuity implies that there exists $\rho_k > 0$ such that $|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \le \rho_k$ implies $|F_n^{(k)} - y_\star|\sqrt{\mathcal{J}_n^{(k)}} < \tau$. In this case, $\gamma_n^{(k),\tau} = \eta(F_n^{(k)} - y_\star)$ and $g_k = \eta(f_\infty^{(k),\tau} - y_\star)$, and therefore

$$|\gamma_n^{(k),\tau} - g_k| = \eta|F_n^{(k)} - f_\infty^{(k),\tau}|.$$

Suppose instead that $|f_\infty^{(k),\tau} - y_\star|\sqrt{J_k^\tau} \geq \tau$. Since $\tau > 0$, we have $f_\infty^{(k),\tau} - y_\star \neq 0$ and $J_k^\tau > 0$. By continuity, there exists $\rho_k > 0$ such that $|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \leq \rho_k$ implies $\operatorname{sgn}(F_n^{(k)} - y_\star) = \operatorname{sgn}(f_\infty^{(k),\tau} - y_\star)$ and $\mathcal{J}_n^{(k)} \geq J_k^\tau/2$. Hence

$$|\gamma_n^{(k),\tau} - g_k| \leq \eta\big||F_n^{(k)} - y_\star| - |f_\infty^{(k),\tau} - y_\star|\big| + \eta\tau\left|(\mathcal{J}_n^{(k)})^{-1/2} - (J_k^\tau)^{-1/2}\right|.$$

Here we used $|\min\{a,b\} - \min\{c,d\}| \leq |a-c| + |b-d|$. Since $x \mapsto x^{-1/2}$ has bounded derivative on $[J_k^\tau/2, \infty)$, there is $L_k < \infty$ such that

$$|\gamma_n^{(k),\tau} - g_k| \leq L_k\left(|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau|\right)$$

whenever $|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \leq \rho_k$.

Combining the two cases, there exist $\rho_k > 0$ and $L_k < \infty$ such that

$$|\gamma_n^{(k),\tau} - g_k| \leq L_k\left(|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau|\right),$$

whenever $|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \leq \rho_k$.

On the event $|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \leq \rho_k$, the local estimate gives

$$\left\|(\gamma_n^{(k),\tau} - g_k)\mathbf{1}_{\{|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| \leq \rho_k\}}\right\|_{L^p} \leq L_k C_k n^{-1/2}.$$

For the complementary event, Markov's inequality gives

$$\mathbb{P}\Big(|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| > \rho_k\Big)^{1/(2p)} \leq \frac{C_k}{\rho_k} n^{-1/2}.$$

By the definition of $\gamma_n^{(k),\tau}$, $|\gamma_n^{(k),\tau}| \leq \eta|F_n^{(k)} - y_\star|$. Hence, Hölder's inequality gives

$$\left\|(\gamma_n^{(k),\tau} - g_k)\mathbf{1}_{\{|F_n^{(k)} - f_\infty^{(k),\tau}| + |\mathcal{J}_n^{(k)} - J_k^\tau| > \rho_k\}}\right\|_{L^p} \leq \left(\eta\|F_n^{(k)} - y_\star\|_{L^{2p}} + |g_k|\right)\frac{C_k}{\rho_k} n^{-1/2}.$$

The assumption $\sup_n \|F_n^{(k)}\|_{L^{2p}} < \infty$ implies $\sup_n \|F_n^{(k)} - y_\star\|_{L^{2p}} < \infty$. Hence, there exists $C_{k,p} < \infty$, independent of n, such that

$$\|\gamma_n^{(k),\tau} - g_k\|_{L^p} \leq C_{k,p} n^{-1/2}.$$

□

G Additional Experiments Details

G.1 Experimental and Training Details

All experiments use the linear residual network

$$h_0(x) = W_0 x, \qquad h_\ell(x) = h_{\ell-1}(x) + \frac{1}{\sqrt{L}} W_\ell h_{\ell-1}(x), \qquad f_{n,L}(x) = \frac{1}{n}\xi^\top h_L(x).$$

At initialization, the entries of W_0, W_ℓ, and ξ are independent centered Gaussians with variances d^{-1}, n^{-1}, and 1, respectively. Only the residual matrices $W_1, \ldots, W_L$ are trained. The input projection and readout remain fixed at initialization. HIGGS uses two independent readout vectors with the same n^{-2} entrywise variance, and their output difference defines the logit margin. Unless stated otherwise, training uses 32-bit floating-point arithmetic, with no momentum, weight decay, learning-rate schedule, or early stopping.

Table 1: **Empirical convergence exponents across training.** Confidence intervals are 95% model-seed bootstrap intervals.

k	$\alpha_n(k)$	95% CI	$\alpha_L(k)$	95% CI
1	0.427	$[0.360, 0.492]$	0.738	$[0.671, 0.807]$
2	0.441	$[0.368, 0.508]$	0.789	$[0.749, 0.837]$
3	0.434	$[0.369, 0.493]$	0.898	$[0.870, 0.929]$
4	0.437	$[0.364, 0.497]$	0.979	$[0.943, 1.017]$
5	0.425	$[0.350, 0.498]$	0.899	$[0.866, 0.936]$
6	0.416	$[0.335, 0.497]$	0.932	$[0.898, 0.971]$
7	0.413	$[0.328, 0.499]$	0.966	$[0.932, 1.006]$
8	0.418	$[0.329, 0.506]$	0.983	$[0.956, 1.013]$
9	0.430	$[0.341, 0.524]$	1.003	$[0.975, 1.035]$
10	0.447	$[0.354, 0.541]$	1.019	$[0.992, 1.052]$
11	0.466	$[0.367, 0.561]$	1.029	$[1.000, 1.063]$
12	0.485	$[0.379, 0.580]$	0.970	$[0.933, 1.006]$
13	0.503	$[0.391, 0.604]$	0.968	$[0.931, 1.005]$
14	0.519	$[0.402, 0.622]$	0.964	$[0.927, 1.002]$
15	0.535	$[0.409, 0.645]$	0.960	$[0.922, 0.998]$

Synthetic regression. We minimize

$$\mathcal{L}(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \big(f_\theta(x_i) - y_i\big)^2$$

by full-batch gradient descent with learning rate 1. We fix a single dataset draw (dataset seed 0) with input dimension $d = 16$, $N = 128$ training examples, and 256 test examples. The input covariance has a geometrically spaced spectrum with condition number 10, and the noiseless targets are normalized by their empirical RMS on the training set. The joint-path experiment uses model seeds $0, \dots, 4$ and records initialization together with 15 gradient updates. The width and depth sweeps use model seeds $0, \dots, 19$. Uncertainty intervals for these sweeps are obtained by resampling model seeds while keeping the dataset fixed.

Empirical rate fits. For Figure 3, the empirical reference $\widehat{f}_\infty^{(k)}$ is the coordinatewise average of the 25 output vectors at the largest common scale $m = 3072$, corresponding to five width–depth scaling paths and five model seeds. For each geometry (n, L), we measure

$$\mathrm{Err}_k(n, L) = \left[\frac{1}{20N} \sum_{s=1}^{20} \sum_{i=1}^{N} \Big(f_{n,L,s}^{(k)}(x_i) - \widehat{f}_\infty^{(k)}(x_i) \Big)^2 \right]^{1/2}.$$

This is the RMS discrepancy of the output vector from the empirical reference, rather than the training loss.

The width sweep fixes $L_{\mathrm{ref}} = 12{,}288$ and uses

$$n \in \{384, 768, 1536, 3072, 6144, 12288\},$$

while the depth sweep fixes $n_{\mathrm{ref}} = 12{,}288$ and uses

$$L \in \{5, 10, 15, 20, 30, 40, 60, 80, 100, 120, 160, 200, 240, 320, 384, 768, 1536, 3072, 6144, 12288\}.$$

The bands in Figure 3(a)–(b) show one standard deviation of the seedwise RMS errors, while the band in Figure 3(c) is obtained from 500 paired model-seed bootstrap refits.

For the width exponent, we fit the slope of $\log \mathrm{Err}_k(n, L_{\mathrm{ref}})$ against $\log n$ over $n \geq 1536$. For the depth exponent, we fit the decreasing regime of the RMS error curve. The large-L error level is estimated from $L \geq 384$ and is used only to determine this fitting range. The regression itself is performed on the raw RMS error.

Table 1 reports the full trajectories in Figure 3(c). The width rate remains close to $n^{-1/2}$ throughout training, whereas the depth-dependent error decays more rapidly and approaches an empirical L^{-1} rate after the initial iterations.

HIGGS. We use a fixed split of 10,500,000 training examples and 500,000 test examples. Each of the 28 features is standardized using the mean and population standard deviation computed on the training split. Writing $g(x)$ for the class-1 versus class-0 logit margin, we minimize the binary logistic loss

$$\mathcal{L}_{\text{HIGGS}} = \frac{1}{N}\sum_{i=1}^{N}\left[\log\left(1+e^{g(x_i)}\right) - y_i g(x_i)\right].$$

We optimize this loss by minibatch SGD with learning rate 1 and batch size 4096. Training runs for 5126 updates, corresponding to two epochs after dropping the final incomplete batch in each epoch. The same two training-set permutations are used for every width, depth, and model seed. We use five model seeds and evaluate the logit margin on a fixed 4096-example test probe. As above, only the residual matrices are updated.

G.2 Additional Results on Width–Depth Commutativity

Synthetic regression. Section 5.1 reports four representative training iterations for one fixed training input. We additionally report the complete trajectory over $k = 0, \ldots, 15$ in Figure 4, using the common scales

$$m \in \{32, 48, 64, 96, 128, 256, 512, 1024, 2048, 3072\},$$

with five model seeds at each scale. We also report four additional fixed training inputs in Figure 5.

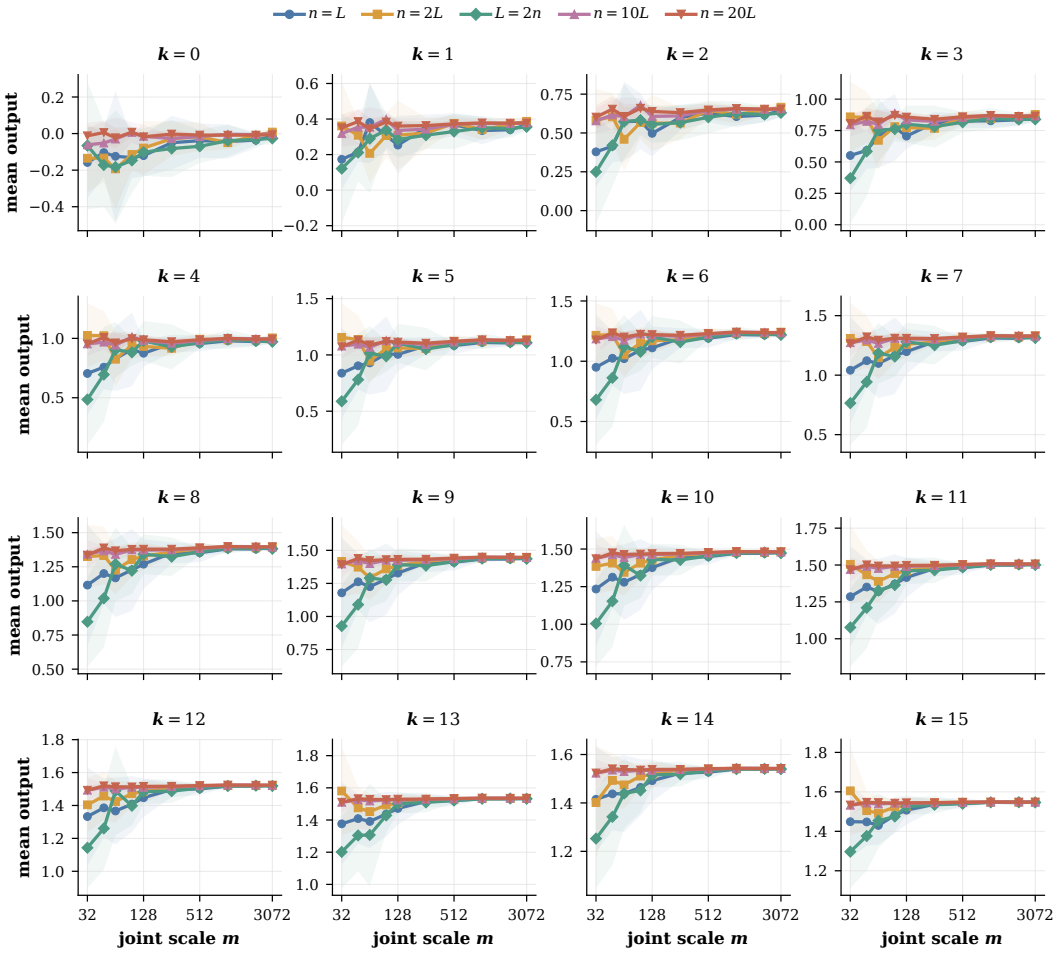


Figure 4: **Synthetic width–depth commutativity across all saved iterations.** Each panel shows the output for training input 86 at one iteration $k = 0, \ldots, 15$. Curves correspond to the five width–depth scaling paths in Section 5.1 and show means over five model seeds, with shaded regions denoting ± 1 standard deviation.

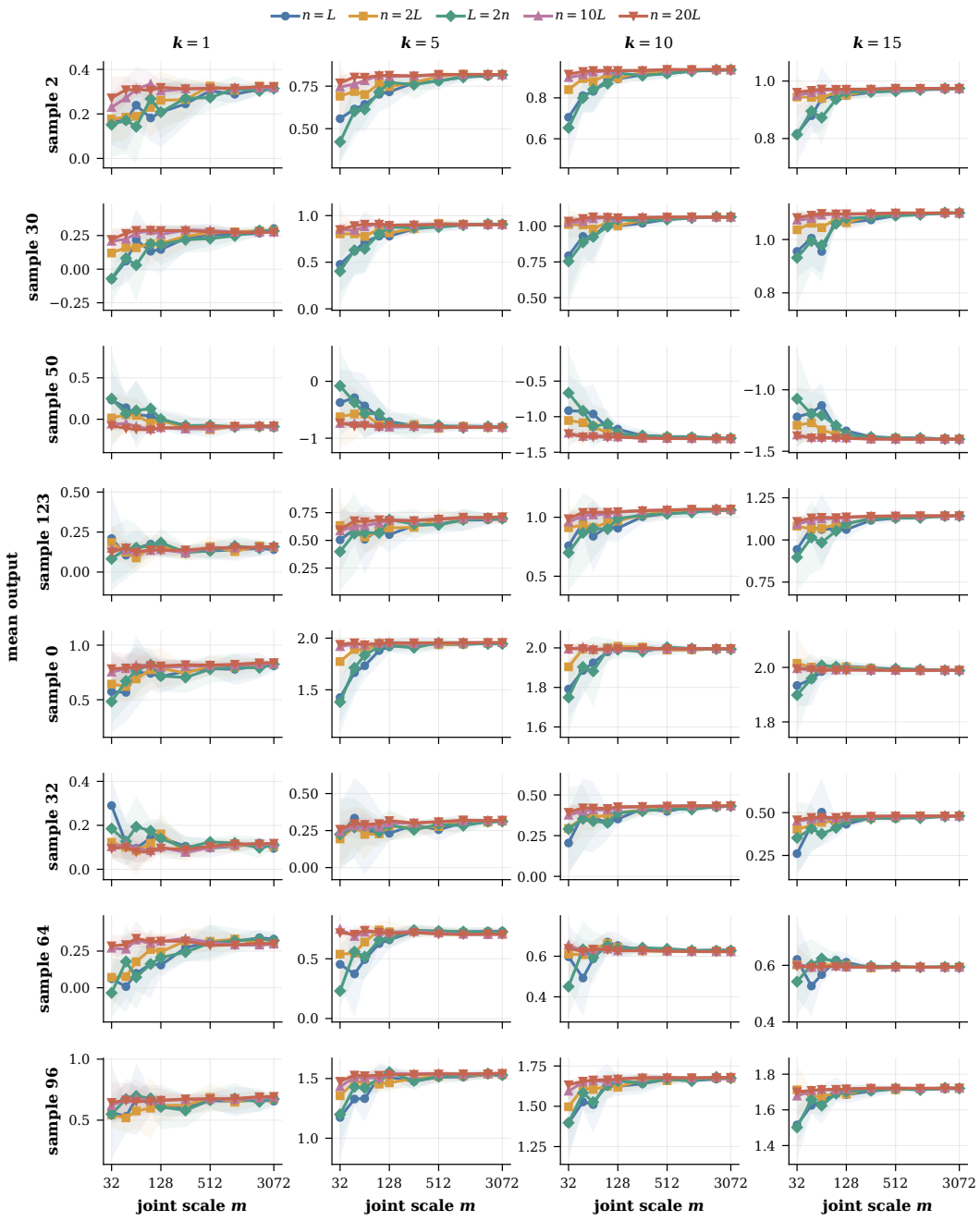


Figure 5: **Synthetic width–depth commutativity at additional training inputs.** Rows correspond to four fixed training inputs, and columns show the representative iterations in Section 5.1. Curves correspond to the same five width–depth scaling paths and show means over five model seeds, with shaded regions denoting ± 1 standard deviation.

HIGGS. Section 5.1 reports four representative training checkpoints for one fixed HIGGS test example. We additionally report all eight saved checkpoints, including initialization, in Figure 6, together with three additional fixed test examples in Figure 7.

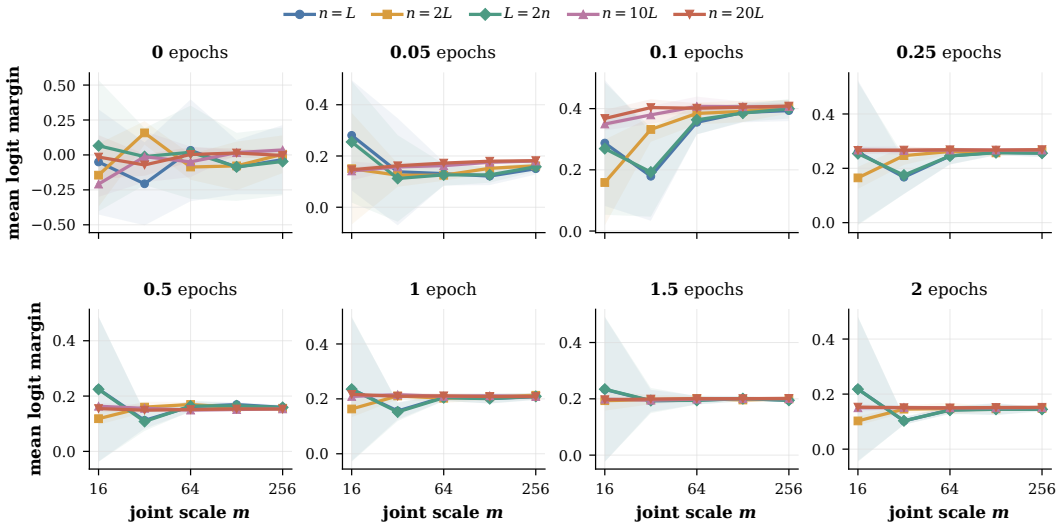


Figure 6: **HIGGS width–depth commutativity across all saved checkpoints.** Each panel shows the class-1 versus class-0 logit margin for the fixed test example used in Section 5.1, from initialization through two training epochs. Curves show means over five model seeds, with shaded regions denoting ± 1 standard deviation, over the common scales $m \in \{16, 32, 64, 128, 256\}$.

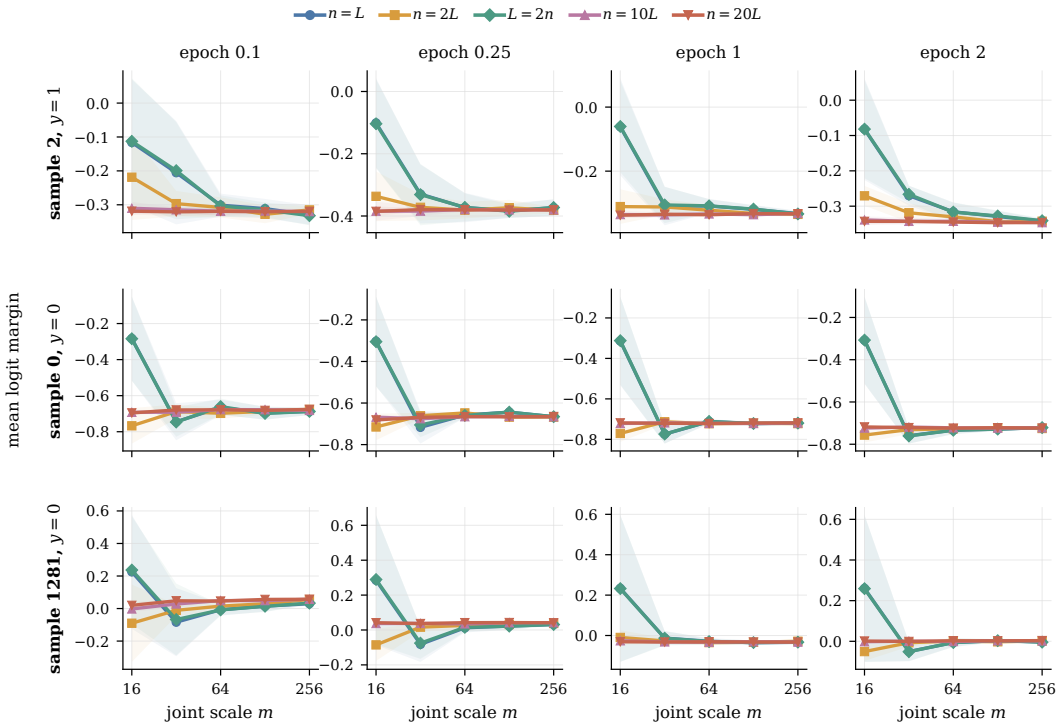


Figure 7: **HIGGS width–depth commutativity at additional test examples.** Rows correspond to three fixed test examples, and columns show the representative checkpoints in Section 5.1. Curves correspond to the same five width–depth scaling paths and show means over five model seeds, with shaded regions denoting ± 1 standard deviation.

G.3 Single-Datum Experiment

The main experiments use a multi-sample regression task. To more closely match the single-datum setting of our theory, we additionally train on a single input–target pair with $q_0 = 1$, $y_\star = 1$, learning rate 1, and five independent model seeds. We evaluate the same five width–depth scaling paths at $m = \min(n, L) \in \{128, 256, 512, 1024, 2048, 3072\}$, for training iterations $k = 0, \ldots, 10$.

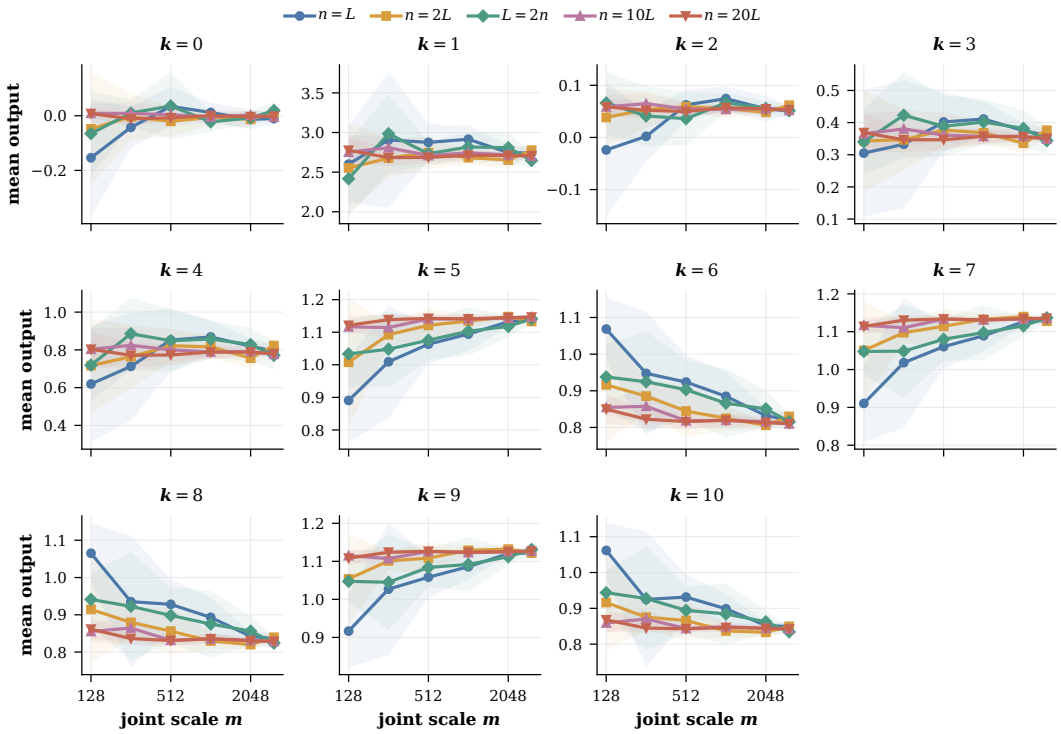


Figure 8: **Width–depth commutativity in the single-datum setting.** Each panel shows the trained scalar output at one training iteration along the five width–depth scaling paths. Curves show the mean over five model seeds, with shaded regions denoting ± 1 standard deviation. The paths exhibit substantial finite-scale differences at small m, but these differences diminish as both width and depth increase.

Figure 8 shows the same transition from finite-scale path dependence to large-scale path independence as in the multi-sample experiment.

G.4 Nonlinear Activation Extension

We further examine whether the finite-scale behavior observed for linear ResNets persists with nonlinear residual blocks. We replace each residual block by

$$h_\ell(x) = h_{\ell-1}(x) + \frac{1}{\sqrt{L}} W_\ell \phi\big(h_{\ell-1}(x)\big),$$

where ϕ is applied elementwise. We consider ReLU, tanh, GELU, SiLU, and sigmoid activations. We use the same synthetic dataset, five width–depth scaling paths, five model seeds, initialization, squared loss, and full-batch gradient descent with learning rate 1 as in Section 5.1. Each model is trained for 10 updates, and we evaluate the common scales

$$m \in \{32, 48, 64, 96, 128, 256\}.$$

The smaller maximum scale reflects the additional memory required to retain the nonlinear hidden states and adjoints during training.

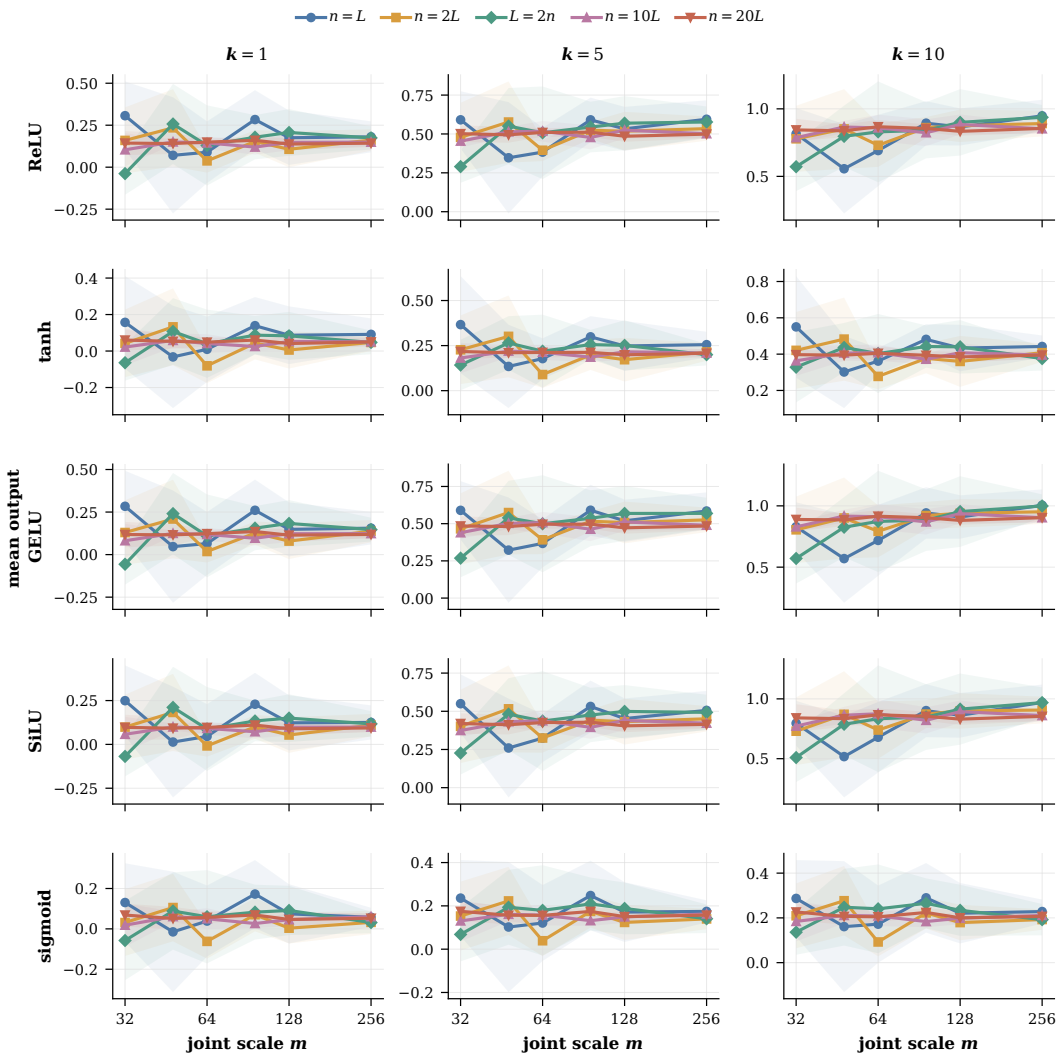


Figure 9: **Width–depth commutativity with nonlinear activations.** Rows correspond to activation functions and columns to training iterations. Each panel shows the output at synthetic training input 86 along the five width–depth scaling paths used in Section 5.1. Curves show means over five model seeds, with shaded regions denoting ± 1 standard deviation.

Across all five activations, the outputs along different width–depth paths become increasingly similar as the common scale grows; see Figure 9. These results provide finite-scale evidence that the commutativity behavior extends beyond the linear architecture, although our current theoretical guarantees do not cover nonlinear activations.

G.5 Computational Details

Lazy Gaussian implementation. The largest synthetic model has $(n, L) = (61{,}440, 3{,}072)$, for which explicitly storing the residual matrices in float32 would require approximately 46.4 TB of memory. We instead use an exact-in-law lazy Gaussian implementation based on Gaussian conditioning and deferred decisions (Bayati & Montanari, 2011), which generates the required matrix–vector products without materializing the residual matrices. This reduces the peak GPU memory usage to 46.9 GB and makes the largest experiments computationally feasible.

Table 2: **Memory usage for the reported experiments.** Dense memory is the float32 storage required to materialize all residual matrices; peak memory is the maximum recorded CUDA allocation.

Experiment	Maximum (n, L)	Backend	Dense memory	Peak memory
Single datum	$(61{,}440, 3{,}072)$	lazy	46.4 TB	31.7 GB
Synthetic commutativity	$(61{,}440, 3{,}072)$	lazy	46.4 TB	46.9 GB
Linear rate sweeps	$(12{,}288, 12{,}288)$	lazy	7.42 TB	25.4 GB
HIGGS commutativity	$(5{,}120, 256)$	dense	26.8 GB	26.9 GB
Nonlinear activations	$(5{,}120, 256)$	dense	26.8 GB	27.6 GB