Feature Learning Dynamics in Infinite-Depth ResNets

Zihan Yao
School of Computing
DePaul University
zyao8@depaul.edu

Ruoyu Wu
Department of Mathematics
Iowa State University
ruoyu@iastate.edu

Tianxiang Gao*
School of Computing
DePaul University
tgao9@depaul.edu

Abstract

Deep neural networks have achieved remarkable success in practice, yet a mechanistic understanding of how features evolve during training remains incomplete, particularly in the large-depth limit. We investigate this question in residual networks (ResNets) with single-layer blocks trained by SGD under depth-μP. Tracking these dynamics is challenging because backpropagation couples forward features and backward gradients via the reuse of weight matrices $\boldsymbol{W}_\ell$ and their transposes $\boldsymbol{W}_\ell^\top$. To isolate this structural dependence, we derive exact conditional Gaussian representations that isolate the weight-reuse effect at finite width and depth. At initialization, its effect vanishes at rate $O(n^{-1/2})$ in coordinatewise L^p uniformly in depth. This allows the joint forward and backward processes to converge to their continuous-depth limits at coordinatewise L^p rates of $O(n^{-1/2} + L^{-1})$ and $O(n^{-1/2} + L^{-1/2})$, respectively, for every $p \geq 2$. After one SGD update, features evolve according to a stochastic differential equation whose drift incorporates both the direct SGD contribution and a persistent response to the initial backward pass. We bound the L^p approximation error for the updated features and network output by $C_{p,L}n^{-1/2} + C_pL^{-1/2}$. However, a linear ResNet counterexample demonstrates that updated feature moments can diverge as depth grows at fixed width, showing that the width and depth limits need not commute. Finally, for multiple updates, we derive mean-field recursions by tracking accumulated feature and gradient histories, recovering the $C_{p,L}n^{-1/2} + C_pL^{-1/2}$ convergence rate under nondegeneracy conditions on the history Gram matrices.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in practice across a wide range of domains [1, 8, 19]. Despite this success, a mechanistic understanding of how useful representations emerge and evolve from data during training remains incomplete [30]. Rigorously characterizing this feature learning process requires tracking how internal feature representations develop across layers and how their interaction with backward error gradients governs the underlying training dynamics.

A fruitful approach is to study the limiting dynamics of simple DNNs trained via stochastic gradient descent (SGD) in the active feature-learning regime [37, 22, 39, 4]. Under suitable large-network scaling limits, this high-dimensional feature-gradient co-evolution can be described by tractable deterministic or stochastic dynamics. These limits are naturally structured along two orthogonal axes: *width* and *depth*. Along the width direction, mean-field theory [22] and Tensor Programs [37] establish principled frameworks for infinite-width feature learning. In particular, maximal update parameterization (μP) ensures that feature updates remain $O(1)$ at infinite width, thereby enabling zero-shot hyperparameter (HP) transfer across network widths [38].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Along the depth axis, residual networks (ResNets) [16] provide the natural setting for analyzing deep feature propagation. Under depth-μP, scaling each residual branch by $1/\sqrt{L}$ supports stable training and hyperparameter transfer across network depths [39, 4]. At initialization, weight matrices are centered and independent across layers. As depth L grows, each residual block makes a small, independent random update to the representations, accumulating into a continuous stochastic differential equation (SDE) indexed by continuous depth $t_\ell = \ell/L \in [0, 1]$. Prior work confirms the validity of this continuous-depth SDE limit at initialization [15, 25] and demonstrates that infinite-width and infinite-depth scaling limits commute for the forward propagation before training [15].

During training, however, backpropagation introduces a structural dependence by coupling weights across layers. Specifically, the forward pass uses weight matrices $\boldsymbol{W}_\ell$, while backpropagation uses their transposes $\boldsymbol{W}_\ell^\top$. This weight reuse couples forward features and backward gradients even at initialization for the first backward pass [33, 35]. Subsequent SGD updates then use both quantities to modify the weight matrices, carrying information from the backward pass into all future forward passes. This dynamic feedback loop gives rise to two fundamental theoretical questions:

> *How can we isolate weight-reuse coupling? How can we quantify its behavior as width and depth grow, both at initialization and during training?*

In this paper, we address these questions in ResNets with single-layer residual blocks trained via SGD under depth-μP. Our central analytical strategy is the construction of exact conditional Gaussian representations that separate the contribution of weight reuse from fresh Gaussian innovations at finite width and depth prior to taking asymptotic limits. These representations enable us to quantify the coupling at initialization and after the first SGD update as width and depth grow. At initialization, weight-reuse coupling in the backward process is a finite-width effect that vanishes as width increases, uniformly in depth. Conversely, after one SGD update, weight reuse contributes a response term to the subsequent forward pass. This term enters the forward feature recursion at the same depth order as the direct SGD drift, surviving to shape the continuous-depth learning dynamics. Our primary contributions are summarized as follows:

- **Gaussian characterization of reused-weight coupling.** We derive exact conditional Gaussian representations that explicitly isolate the contribution of weight reuse from fresh Gaussian innovations for any finite SGD updates. These representations pave the foundation for our asymptotic analysis, which we deploy at initialization and following a one-step SGD update.
- **Initialization-time gradient independence with a depth-uniform rate.** We compare the true backward process with an auxiliary process that uses decoupled backward weights in backpropagation, as in the gradient independence assumption (GIA) [33, 35]. The coordinatewise L^p error between these processes is $O(n^{-1/2})$, uniformly in depth. This estimate allows us to establish a joint continuous-depth SDE limit for the forward and backward processes. The approximation error is $O(n^{-1/2} + L^{-1/2})$ in L^p along arbitrary sequences with $n, L \to \infty$.
- **Quantitative feature dynamics after one SGD update.** We derive a closed continuous-depth SDE limit for the updated forward features that retains the response induced by weight reuse. We first approximate the updated network by its infinite-width dynamics at fixed depth, then control the continuous-depth approximation of those dynamics. Together, these estimates give an L^p error bound of $C_{p,L} n^{-1/2} + C_p L^{-1/2}$ for the updated features and network output.
- **Failure of depth-uniform width convergence during training.** By constructing a counterexample using linear ResNet, we show that an updated feature coordinate's L^p norm diverges as $L \to \infty$ at fixed width n, while the infinite-width norm remains bounded uniformly in depth. This result rigorously refutes depth-uniform width convergence post-initialization, proving that width and depth limits under depth-μP fail to commute during active feature learning.

Our main results concern a single SGD update. In Section 5 and Appendix F, we also explore how the analysis extends to multiple updates and derive candidate continuous-depth mean-field equations. Nearly singular history Gram matrices complicate width convergence, and the dependence of activation derivatives on backward noise complicates depth convergence. With nondegenerate history Gram matrices, we recover the $O(n^{-1/2})$ width convergence rate at fixed depth. For linear activation and a depth-uniform lower bound on the older Gram blocks, we justify the continuous-depth equations and recover the $O(L^{-1/2})$ depth convergence rate.

2 Related Work

Width limits and feature learning. An important line of theoretical work studies the large-width limit of neural networks. In the Neural Tangent Kernel (NTK) regime [18, 20, 3], training is well approximated by kernel regression with an essentially fixed kernel, leading to strong optimization and generalization guarantees for overparameterized networks [2, 9, 40, 13, 14]. However, this lazy-training regime [7] does not capture substantial feature evolution. Mean-field parameterizations provide an infinite-width feature-learning alternative [22, 6, 31], with convergence guarantees for shallow networks [24, 26], but standard mean-field scalings become degenerate in deep architectures, leading to vanishing signals or gradients [10, 23]. Tensor Programs and μP [34, 35, 36, 37, 38] clarify which width scalings yield nontrivial feature learning and enable HP transfer.

Depth limits and training dynamics. Large-depth analyses study how signals and gradients behave as depth increases, often in conjunction with width. Classical signal-propagation work identified stability conditions such as the edge of chaos [27, 29]. For ResNets, scaling the residual branch by $1/\sqrt{L}$ restores stable propagation and yields an SDE view of forward feature propagation at initialization in joint width-depth limits [15]; related large-depth stability studies also appear in implicit models and Neural ODEs [12, 13, 11]. More recent work shows that depth-μP scaling enables depthwise HP transfer [39, 4] and that feature learning can persist in infinite-depth networks [39]. These results motivate an SDE view introduced in [4] to describe large-depth training dynamics through feature and gradient covariances and response functions.

Gradient independence and weight reuse. The gradient independence assumption replaces backpropagation weights with independent copies of forward weights, yielding correct infinite-width limits in known regimes [33, 35]. We analyze weight reuse in finite networks via exact conditional Gaussian representations. At initialization, this yields a depth-uniform rate for the approximation, extending joint width–depth analysis to the backward process. Post-SGD, it isolates a persistent response term in the continuous-depth dynamics, and a linear counterexample shows the depth-first L^p limit can fail post-update despite a valid width-first limit.

3 Preliminaries

Model and Training Setup. In this paper, we consider a simple one-layer ResNet under depth-μP scaling [39], which maps an input $\boldsymbol{x} \in \mathbb{R}^d$ through the scaled residual stream defined as follows

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{n}\,\boldsymbol{v}^\top \boldsymbol{h}_L, \quad \boldsymbol{h}_\ell = \boldsymbol{h}_{\ell-1} + \frac{1}{\sqrt{Ln}}\,\boldsymbol{W}_\ell\,\phi(\boldsymbol{h}_{\ell-1}), \quad \boldsymbol{h}_0 = \frac{1}{\sqrt{d}}\,\boldsymbol{U}\boldsymbol{x}, \quad \forall \ell \in [L], \tag{1}$$

where ϕ is an activation function, and $\boldsymbol{U} \in \mathbb{R}^{n\times d}$, $\boldsymbol{W}_\ell \in \mathbb{R}^{n\times n}$, and $\boldsymbol{v} \in \mathbb{R}^n$ are trainable parameters. We denote the collection of parameters by $\boldsymbol{\theta} := (\boldsymbol{v}, \{\boldsymbol{W}_\ell\}_{\ell=1}^L, \boldsymbol{U})$ and initialize all entries independently $\boldsymbol{\theta}_i \sim \mathcal{N}(0,1)$. Here, the $1/\sqrt{n}$ width scaling and $1/\sqrt{L}$ depth scaling on the residual branches follow directly from the standard μP [37] and depth-μP [39], respectively.

Given a loss $\mathcal{L}$, we train the ResNet with stream SGD using one sample $(\boldsymbol{x}^k, y^k)$ per iteration:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(f^k, y^k) = \boldsymbol{\theta}^k - \eta \chi^k \nabla_{\boldsymbol{\theta}^k} f^k, \tag{2}$$

where $f^k := f(\boldsymbol{x}^k;\boldsymbol{\theta}^k)$ and $\chi^k := \partial_f \mathcal{L}(f^k, y^k)$. To obtain nontrivial feature updates in the infinite-width and infinite-depth limits, following μP and depth-μP, we use the learning rate $\eta = \eta_c n$, where $\eta_c > 0$ is the fixed effective learning rate, independent of the network width and depth.

Pre- vs. Post-activation ResNets. While the original ResNet [16] used the post-activation design Eq. (27), modern architectures commonly adopt the pre-activation form Eq. (1) [17], including Transformers [28, 5]. Beyond empirical practice, the following proposition shows that post-activation ResNets can exhibit hidden-state divergence at large depth; see Appendix B for the proof.

Proposition 3.1. *Let ϕ satisfy the following **positive dominance** condition: there exist nonnegative constants c_1, c_2, not both zero, such that $\mathbb{E}[\phi(wx)] \geq c_1|x| + c_2$, $\forall x \in \mathbb{R}$, where $w \sim \mathcal{N}(0,1)$. Then, in the post-activation ResNet Eq. (27), the expected hidden state obeys, for every $i \in [n]$,*

$$\mathbb{E}[\boldsymbol{h}_{L,i}] \geq c_1 \left(1 + \frac{c_1}{\sqrt{Ln}}\right)^L \frac{\|\boldsymbol{x}\|}{\sqrt{d}} + c_2\sqrt{L}.$$

For activations such as ReLU, where $c_1 = 1/\sqrt{2\pi}$ and $c_2 = 0$, Proposition 3.1 shows that the hidden states $\boldsymbol{h}_\ell$ may diverge as depth grows, even under the residual scaling $1/\sqrt{L}$. Figure 4 illustrates this behavior. The superior stability of the pre-activation design justifies focusing our subsequent analysis on Eq. (1); it also avoids additional normalization or carefully balanced depth-width scaling assumptions often needed in analyses of post-activation ResNets [25, 39, 21].

4 Gradient Independence and Joint Width–Depth Limits at Initialization

We first analyze the weight-reuse effect at initialization. By using a conditional Gaussian representation, we clearly isolate this effect from residual Gaussian innovation at finite width and depth, and then establish a joint continuous-depth limit for the forward features and backward gradients.

Weight reuse in backpropagation. The backward process starts from the readout $\boldsymbol{v}$ and satisfies

$$\boldsymbol{g}_L = \boldsymbol{v}, \qquad \boldsymbol{g}_{\ell-1} = \boldsymbol{g}_\ell + \frac{1}{\sqrt{Ln}}\,\phi'(\boldsymbol{h}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell, \qquad \ell \in [L], \tag{3}$$

where the backward gradient $\boldsymbol{g}_\ell := n\nabla_{\boldsymbol{h}_\ell} f$ and $\odot$ denotes elementwise multiplication. The weight-reuse effect coupling occurs since the forward pass Eq. (1) uses $\boldsymbol{W}_\ell$, while the backward pass Eq. (3) uses its transpose. To measure the effect of this reuse, we compare $\boldsymbol{g}$ with an auxiliary backward process $\widetilde{\boldsymbol{g}}$. The auxiliary process starts from the same $\boldsymbol{v}$ and uses the same forward features in Eq. (3), with $\boldsymbol{W}_\ell^\top$ replaced by $\widetilde{\boldsymbol{W}}_\ell^\top$. The matrices $\widetilde{\boldsymbol{W}}_\ell$ are independent of the forward trajectory. This is the gradient independence assumption (GIA) [33, 35]. The appendix gives the joint construction used to compare the two processes in Eq. (59) and Eq. (60).

Gaussian representation. Write $\boldsymbol{\phi}_\ell := \phi(\boldsymbol{h}_\ell)$ and define the forward and backward weight queries

$$\boldsymbol{a}_\ell := \frac{1}{\sqrt{n}}\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}, \qquad \boldsymbol{b}_\ell := \frac{1}{\sqrt{n}}\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell. \tag{4}$$

The forward and backward recursions then take the form

$$\boldsymbol{h}_\ell = \boldsymbol{h}_{\ell-1} + \frac{1}{\sqrt{L}}\boldsymbol{a}_\ell, \qquad \boldsymbol{g}_{\ell-1} = \boldsymbol{g}_\ell + \frac{1}{\sqrt{L}}\boldsymbol{b}_\ell \odot \phi'(\boldsymbol{h}_{\ell-1}). \tag{5}$$

At initialization, $\boldsymbol{W}_\ell$ has i.i.d. standard Gaussian entries and is independent of $\boldsymbol{\phi}_{\ell-1}$. Conditional on $\boldsymbol{\phi}_{\ell-1}$, the forward query $\boldsymbol{a}_\ell$ is therefore centered Gaussian with covariance $\|\boldsymbol{\phi}_{\ell-1}\|^2 \boldsymbol{I}_n/n$. When the backward pass reaches layer ℓ, the forward query has already revealed the action of $\boldsymbol{W}_\ell$ on $\boldsymbol{\phi}_{\ell-1}$. The following proposition shows how this information enters the backward query at finite width and depth. Lemma C.1 in Appendix C gives the general conditional representation. Corollary C.3 specializes it to initialization, with the forward and backward laws given in Eq. (41) and Eq. (42).

Proposition 4.1 (Gaussian representation at initialization)**.** *Let $\boldsymbol{P}_{\ell-1}$ denote the orthogonal projection onto the span of $\boldsymbol{\phi}_{\ell-1}$. The forward and backward processes admit a joint realization in which*

$$\boldsymbol{a}_\ell = \frac{\|\boldsymbol{\phi}_{\ell-1}\|}{\sqrt{n}}\,\boldsymbol{z}_\ell, \qquad \boldsymbol{b}_\ell = \frac{\langle \boldsymbol{a}_\ell, \boldsymbol{g}_\ell\rangle}{\|\boldsymbol{\phi}_{\ell-1}\|^2}\,\boldsymbol{\phi}_{\ell-1} + \frac{\|\boldsymbol{g}_\ell\|}{\sqrt{n}}(\boldsymbol{I}_n - \boldsymbol{P}_{\ell-1})\widetilde{\boldsymbol{z}}_\ell. \tag{6}$$

Here $\boldsymbol{z}_\ell$ and $\widetilde{\boldsymbol{z}}_\ell$ are standard Gaussian innovations introduced in computation order, as in Lemma C.1. If $\boldsymbol{\phi}_{\ell-1} = 0$, we set $\boldsymbol{P}_{\ell-1} = 0$ and interpret the aligned term as zero.

The aligned term in Eq. (6) follows from $\langle \boldsymbol{b}_\ell, \boldsymbol{\phi}_{\ell-1}\rangle = \langle \boldsymbol{g}_\ell, \boldsymbol{a}_\ell\rangle$. Both $\boldsymbol{g}_\ell$ and $\boldsymbol{a}_\ell$ are known before the backward query $\boldsymbol{b}_\ell$, so this identity fixes the projection of $\boldsymbol{b}_\ell$ onto the forward feature direction.

The auxiliary backward process uses an independent copy of the layer weights in the backward pass and retains the same activation derivatives $\phi'(\boldsymbol{h}_{\ell-1})$ from the forward pass. Since the auxiliary matrix at layer ℓ is independent of the history before its backward query, the conditional law is isotropic:

$$\widetilde{\boldsymbol{b}}_\ell \overset{d}{=} \frac{\|\widetilde{\boldsymbol{g}}_\ell\|}{\sqrt{n}}\,\widetilde{\boldsymbol{z}}_\ell, \qquad \widetilde{\boldsymbol{z}}_\ell \sim \mathcal{N}(0, \boldsymbol{I}_n). \tag{7}$$

We can use the same Gaussian innovation $\widetilde{\boldsymbol{z}}_\ell$ in both representations. In this coupling, the true query uses the component along $\boldsymbol{\phi}_{\ell-1}$, whereas the auxiliary query samples that component along with the orthogonal fluctuations. This gives a direct way to compare the two backward recursions while keeping their dependence on the forward features through the activation derivatives.

Depth-uniform gradient independence. The Gaussian representation above isolates the effect of weight reuse at a single layer. The next result bounds how much this effect accumulates over the entire backward pass, uniformly in the network depth.

Proposition 4.2 (Depth-uniform backward decoupling)**.** *Suppose $q_0 > 0$ and ϕ is differentiable, globally Lipschitz, and not identically zero. For every $p \geq 2$, the true and auxiliary backward processes admit a coupling such that*

$$\sup_{L\geq 1}\sup_{0\leq \ell\leq L}\sup_{1\leq i\leq n}\left\|g_{\ell,i}^{n,L}-\widetilde{g}_{\ell,i}^{n,L}\right\|_{L^p}\leq C_p n^{-1/2}, \tag{8}$$

where C_p is independent of n and L.

Proposition D.4 in Appendix D proves Eq. (8) using the matrix coupling in Eq. (59). The auxiliary backward recursion uses independent copies of the weights and retains the activation derivatives $\phi'(\boldsymbol{h}_{\ell-1})$ from the original forward pass. These derivatives carry the dependence on the forward trajectory into the limiting backward equation.

Joint width–depth limit. The depth-uniform estimate lets us control the width and depth errors simultaneously. The forward and backward processes have different depth rates, as stated below.

Theorem 4.3 (Forward–backward SDE limit at initialization)**.** *Suppose $q_0 > 0$ and $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. Let $H_0 \sim \mathcal{N}(0, q_0)$, $V \sim \mathcal{N}(0,1)$, and standard Brownian motions B and $\widetilde{B}$ be mutually independent. The system*

$$H_t = H_0 + \int_0^t \sqrt{\mathbb{E}[\phi(H_s)^2]}\,\mathrm{d}B_s, \quad G_t = V + \int_t^1 \phi'(H_s)\sqrt{\mathbb{E}[G_s^2]}\,\mathrm{d}\widetilde{B}_s \tag{9}$$

has a unique continuous solution with finite second moments of the path supremum. For every $p \geq 2$ and every n, L, there is a joint coupling of the finite-network states and independent coordinate copies $(H_{\cdot,i}, G_{\cdot,i})$ of this solution such that, with $t_\ell = \ell/L$,

$$\begin{aligned}\sup_{\ell,i}\left\|h_{\ell,i}^{n,L}-H_{t_\ell,i}\right\|_{L^p}&\leq C_p\big(n^{-1/2}+L^{-1}\big),\\ \sup_{\ell,i}\left\|g_{\ell,i}^{n,L}-G_{t_\ell,i}\right\|_{L^p}&\leq C_p\big(n^{-1/2}+L^{-1/2}\big).\end{aligned} \tag{10}$$

The suprema range over $0 \leq \ell \leq L$ and $1 \leq i \leq n$, and C_p is independent of n and L.

Together, Proposition 4.2 and Eq. (10) quantify the effect of weight reuse at initialization. Its effect on backward propagation is $O(n^{-1/2})$ in coordinatewise L^p, uniformly in depth. The forward and backward processes converge to their continuous-depth SDE limit at coordinatewise L^p rates of $O(n^{-1/2} + L^{-1})$ and $O(n^{-1/2} + L^{-1/2})$, respectively. The same limit therefore holds along any sequence with $n, L \to \infty$. In this sense, the width and depth limits commute at initialization. The initialization panels of Figure 1 illustrate this convergence. The next section examines how the first SGD update changes the effect of weight reuse.

5 Forward–Backward Coupling After One SGD Update

At initialization, the effect of reusing weights in backpropagation vanishes as width grows, uniformly in depth (Proposition 4.2). An SGD update carries information from the backward pass into the next forward pass. We identify the resulting coupling at finite width and depth and show how it enters the limiting dynamics of the updated features.

The updated forward query. We restore the superscripts 0 and 1 to distinguish the initial and updated networks. We train on $(\boldsymbol{x}^0, y)$ with learning rate $\eta = \eta_c n$ and evaluate the updated network at $\boldsymbol{x}^1$. Write $\chi^0 = \partial_f \mathcal{L}(f^0, y)$, $\boldsymbol{\phi}_\ell^j = \phi(\boldsymbol{h}_\ell^j)$, and $q_x^{r,s} = \langle \boldsymbol{x}^r, \boldsymbol{x}^s\rangle/d$. For $r, s \in \{0,1\}$, define the empirical feature kernel $C_{\phi,\ell}^{r,s} = n^{-1}\langle \boldsymbol{\phi}_\ell^r, \boldsymbol{\phi}_\ell^s\rangle$. Substituting the parameter update Eq. (96) from Appendix E into the forward recursion gives

$$\boldsymbol{h}_\ell^1 = \boldsymbol{h}_{\ell-1}^1 + \frac{1}{\sqrt{L}}\boldsymbol{a}_\ell^1 - \frac{\eta_c \chi^0}{L}\boldsymbol{C}_{\phi,\ell-1}^{1,0}\boldsymbol{g}_\ell^0, \qquad \boldsymbol{a}_\ell^1 := \frac{1}{\sqrt{n}}\boldsymbol{W}_\ell^0\boldsymbol{\phi}_{\ell-1}^1. \tag{11}$$

The last term is the direct contribution of the layer's SGD update. The query $\boldsymbol{a}_\ell^1$ uses the initial matrix $\boldsymbol{W}_\ell^0$ on the updated features, so it can carry information from the initial backward pass as well.

Proposition 5.1 (Gaussian representation after one SGD update)**.** *Let P_u be the orthogonal projection onto the span of u, with $P_0 = 0$, and set $\boldsymbol{\phi}^1_{\ell-1,\perp} = (\boldsymbol{I}_n - P_{\boldsymbol{\phi}^0_{\ell-1}})\boldsymbol{\phi}^1_{\ell-1}$. Conditional on the computation history before evaluating $\boldsymbol{a}^1_\ell$,*

$$\boldsymbol{a}^1_\ell \overset{d}{=} \frac{C^{1,0}_{\phi,\ell-1}}{C^{0,0}_{\phi,\ell-1}}\boldsymbol{a}^0_\ell + \frac{\langle \boldsymbol{b}^0_\ell, \boldsymbol{\phi}^1_{\ell-1,\perp}\rangle}{\|\boldsymbol{g}^0_\ell\|^2}\boldsymbol{g}^0_\ell + \frac{\|\boldsymbol{\phi}^1_{\ell-1,\perp}\|}{\sqrt{n}}(\boldsymbol{I}_n - P_{\boldsymbol{g}^0_\ell})\boldsymbol{z}^1_\ell, \quad \boldsymbol{z}^1_\ell \sim \mathcal{N}(0, \boldsymbol{I}_n) \tag{12}$$

where $\boldsymbol{z}^1_\ell$ is independent of that history. Ratios with zero denominators are interpreted as zero.

This is the forward query in Proposition 5.1 derived from Lemma C.1 in Appendix C. The first term of Eq. (12) reuses the first forward query $\boldsymbol{a}^0_\ell$ on the part of the updated features aligned with the initial features. On the orthogonal part, the first backward query $\boldsymbol{b}^0_\ell$ fixes the component along $\boldsymbol{g}^0_\ell$ through the overlap $\langle \boldsymbol{b}^0_\ell, \boldsymbol{\phi}^1_{\ell-1,\perp}\rangle$. The remaining component is a fresh Gaussian fluctuation. This overlap is the source of the response in the width limit.

The fixed-depth mean field. Let $\Delta = L^{-1}$ and $\lambda = \eta_c \partial_f \mathcal{L}(0, y)$, using the zero initialization output in the width limit, as stated in Corollary D.7. We write the mean-field equations for a representative coordinate and set

$$\bar{C}^{r,s,L}_{\phi,\ell} = \mathbb{E}[\phi(\bar{h}^{r,L}_\ell)\phi(\bar{h}^{s,L}_\ell)], \qquad r^L_\ell = \mathbb{E}[(\bar{g}^{0,L}_\ell)^2].$$

Here $(\bar{h}^{0,L}, \bar{g}^{0,L})$ are the initialization mean fields in Eq. (49) and Eq. (50). The updated forward mean field satisfies

$$\begin{aligned} \bar{h}^{1,L}_0 &= \xi^1_0 - \lambda q^{1,0}_x \bar{g}^{0,L}_0, \\ \bar{h}^{1,L}_\ell &= \bar{h}^{1,L}_{\ell-1} + \sqrt{\Delta}\,\xi^{1,L}_\ell + \Delta\big(\bar{R}^{1,0,L}_\ell - \lambda \bar{C}^{1,0,L}_{\phi,\ell-1}\big)\bar{g}^{0,L}_\ell. \end{aligned} \tag{13}$$

The input pair (ξ^0_0, ξ^1_0) is centered Gaussian with covariance $q^{r,s}_x$. The forward innovations $(\xi^{0,L}_\ell, \xi^{1,L}_\ell)$ are centered Gaussian pairs, independent across layers and of the initialization backward innovations, with covariance entries $\bar{C}^{r,s,L}_{\phi,\ell-1}$. The response coefficient is

$$\bar{R}^{1,0,L}_\ell = \frac{\mathbb{E}[\zeta^{0,L}_\ell \phi(\bar{h}^{1,L}_{\ell-1})]}{\sqrt{\Delta}\, r^L_\ell}, \tag{14}$$

where $\zeta^{0,L}_\ell$ is the Gaussian innovation in the initialization backward mean field. Equation Eq. (13) thus separates the response drift $\bar{R}^{1,0,L}_\ell$ from the direct SGD drift $-\lambda \bar{C}^{1,0,L}_{\phi,\ell-1}$. The appendix gives the innovation construction in Eq. (103) and the response formula in Eq. (105).

Proposition 5.2 (Width convergence of the updated forward pass)**.** *Fix L, deterministic inputs $\boldsymbol{x}^0 \neq 0$ and $\boldsymbol{x}^1$, a label y, and $\eta_c \geq 0$. Suppose $\phi \in C^1(\mathbb{R})$ is not identically zero, ϕ and ϕ' are globally Lipschitz, and $u \mapsto \partial_f \mathcal{L}(u, y)$ is globally Lipschitz. For every $p \geq 2$, there is a coupling with independent coordinate copies of Eq. (13) such that*

$$\max_{0\leq \ell \leq L} \sup_{1\leq i\leq n} \|h^{1,n,L}_{\ell,i} - \bar{h}^{1,L}_{\ell,i}\|_{L^p} + \|f^{1,n,L} - \bar{f}^{1,L}\|_{L^p} \leq C_{p,L} n^{-1/2}. \tag{15}$$

Here $\bar{f}^{1,L} = \mathbb{E}[(v^0 - \lambda \bar{h}^{0,L}_L)\bar{h}^{1,L}_L]$, and $C_{p,L}$ is independent of n.

Proposition 5.2 combines the feature bound from Proposition E.4 with the output bound Eq. (137) from Corollary E.5. The feature estimate also controls the full forward weight query, including its Gaussian and response contributions, in Eq. (132).

The one-update width estimate relies on the lower bounds for the initial activation and gradient second moments, $\bar{C}^{0,0,L}_{\phi,\ell-1}$ and $\bar{C}^{0,0,L}_{g,\ell}$, established in Lemma D.1, which keep the limiting scalar regression denominators away from zero. Over k SGD updates, the corresponding regressions involve the history Gram matrices $\bar{C}^{\leq k,\leq k}_{\phi,\ell-1}$ and $\bar{C}^{\leq k,\leq k}_{g,\ell}$. If their smallest eigenvalues are bounded away from zero across layers, Proposition F.4 recovers the $O(n^{-1/2})$ width convergence rate for fixed L and k. However, these history matrices can become singular when successive directions are linearly dependent. Figure 3 illustrates this near-degeneracy in finite networks, where the empirical matrices exhibit small minimum eigenvalues.

Continuous-depth dynamics. To pass to continuous depth in Eq. (13), we first express the response $\bar{R}_\ell^{1,0,L}$ in terms of the forward states and the accumulated coefficient $\alpha_{\ell-1}^L$. Using Gaussian integration by parts, Lemma E.9 gives

$$\bar{R}_\ell^{1,0,L} = \alpha_{\ell-1}^L \mathbb{E}[\phi'(\bar{h}_{\ell-1}^{1,L})\phi'(\bar{h}_{\ell-1}^{0,L})]. \tag{16}$$

Starting from $\alpha_0^L = -\lambda q_x^{1,0}$, the recursion in Eq. (141) updates α_ℓ^L via the response and direct SGD increments. Because this same increment multiplies $\bar{g}_\ell^{0,L}$ in the feature drift, tracking α^L alongside the updated features successfully closes the recursion.

With the response closed, we reformulate the updated forward recursion on the mesh $t_\ell = \ell/L$ using Brownian increments. To match the exact cross-covariance with the initial pass while supplying the correct residual variance, we share the Brownian motion B^0 and introduce an independent Brownian motion B^1. Writing $\bar{C}_{\phi,\ell-1}^{r,s,L} = \mathbb{E}[\phi(\bar{h}_{\ell-1}^{r,L})\phi(\bar{h}_{\ell-1}^{s,L})]$, we set

$$\sigma_{0,\ell-1}^L = \frac{\bar{C}_{\phi,\ell-1}^{1,0,L}}{\sqrt{\bar{C}_{\phi,\ell-1}^{0,0,L}}}, \qquad \sigma_{1,\ell-1}^L = \sqrt{\bar{C}_{\phi,\ell-1}^{1,1,L} - (\sigma_{0,\ell-1}^L)^2}.$$

From initial state $\bar{h}_0^{1,L} = \xi_0^1 + \alpha_0^L \bar{g}_0^{0,L}$, the discrete path Eq. (13) satisfies

$$\bar{h}_\ell^{1,L} - \bar{h}_{\ell-1}^{1,L} = \sum_{j=0}^{1} \sigma_{j,\ell-1}^L (B_{t_\ell}^j - B_{t_{\ell-1}}^j) + (\alpha_\ell^L - \alpha_{\ell-1}^L)\bar{g}_\ell^{0,L}.$$

We keep the same Brownian motions as $L \to \infty$. Letting (H^0, G^0) denote the initialization limit from Theorem 4.3, the initial state converges to $H_0^1 = \xi_0^1 - \lambda q_x^{1,0} G_0^0$. Replacing discrete sums with integrals yields the limiting continuous-depth system:

$$\begin{aligned} \alpha(t) &= -\lambda q_x^{1,0} + \int_0^t [R^{1,0}(s) - \lambda C_\phi^{1,0}(s)]\,\mathrm{d}s, \qquad R^{1,0}(t) = \alpha(t) C_{\phi'}^{1,0}(t), \\ H_t^1 &= \xi_0^1 - \lambda q_x^{1,0} G_0^0 + \sum_{j=0}^{1} \int_0^t \sigma_j(s)\,\mathrm{d}B_s^j + \int_0^t [R^{1,0}(s) - \lambda C_\phi^{1,0}(s)] G_s^0\,\mathrm{d}s. \end{aligned} \tag{17}$$

The system Eq. (17) defines a fixed-point problem because its coefficients depend on moments of the unknown process H^1 and its joint moments with H^0. The Lipschitz assumptions on ϕ and ϕ' control the drift kernels. Lemma D.8 bounds $C_\phi^{0,0}(t) = F(q(t))$ away from zero uniformly in t. With this lower bound, Lemma E.10 gives Lipschitz control of both diffusion coefficients on bounded sets, including when σ_1 vanishes. Proposition E.12 uses these estimates to obtain a contraction on short depth intervals and uses moment bounds to extend the solution to $[0, 1]$.

The same coefficient estimates control the discretization error. On the common Brownian coupling, the Burkholder–Davis–Gundy inequality bounds the stochastic error in terms of the differences between the diffusion coefficients. The initialization estimates Eq. (91) and Eq. (85) control the error from replacing $\bar{g}_\ell^{0,L}$ by G_s^0 in the drift on each grid interval. Proposition E.13 combines these bounds and applies Gronwall's inequality to obtain the $O(L^{-1/2})$ depth rate.

Theorem 5.3 (Continuous-depth limit of the updated forward pass)**.** *Under the assumptions of Proposition 5.2, Eq. (17) has a unique strong solution (α, H^1) on $[0, 1]$, with α deterministic. For every $p \geq 2$, there is a coupling such that*

$$\max_{0 \leq \ell \leq L} \|\bar{h}_\ell^{1,L} - H_{t_\ell}^1\|_{L^p} \leq C_p L^{-1/2}. \tag{18}$$

Combining the depth estimates with Eq. (15), as in Corollary E.16, yields

$$\max_{0 \leq \ell \leq L} \sup_{1 \leq i \leq n} \|h_{\ell,i}^{1,n,L} - H_{t_\ell,i}^1\|_{L^p} + \|f^{1,n,L} - f_\infty^1\|_{L^p} \leq C_{p,L} n^{-1/2} + C_p L^{-1/2}, \tag{19}$$

where $H_{\cdot,i}^1$ are independent copies of H^1 and $f_\infty^1 = \mathbb{E}[(v^0 - \lambda H_1^0) H_1^1]$.

The response $R^{1,0}(t) G_t^0$ enters the limiting drift at the same order as the direct SGD term. For example, if $\phi = \tanh$ and $\lambda q_x^{1,0} \neq 0$, then $R^{1,0}(0) = -\lambda q_x^{1,0} C_{\phi'}^{1,0}(0) \neq 0$. By continuity, the response coefficient remains nonzero over an interval of depth.

Failure of depth-uniform width convergence. The width constant $C_{p,L}$ in Eq. (19) can depend on depth. The next example shows why this dependence matters.

Proposition 5.4 (A depth-first moment obstruction)**.** *Consider Gaussian initialization, the linear activation $\phi(u) = u$, the squared loss $\mathcal{L}(f, 0) = f^2/2$, and $d = 1$ with $\boldsymbol{x}^0 = \boldsymbol{x}^1 = 1$. After one SGD update with $\eta = \eta_c n$ and $\eta_c > 0$, for every fixed $n \geq 1$, $p \geq 2$, and $1 \leq i \leq n$,*

$$\lim_{L\to\infty} \|h_{L,i}^{1,n,L}\|_{L^p} = \infty, \qquad \sup_{L\geq 1} \|\bar{h}_{L,i}^{1,L}\|_{L^p} \leq C_p. \tag{20}$$

In this example, the initial output is zero in the width limit, so the mean-field update vanishes. At finite width, rare initializations produce growth across the updated layers large enough to make the moments diverge. Thus the width-first limit exists in L^p, while the depth-first limit fails in L^p. Proposition E.7 in Appendix E.3 proves Eq. (20).

Beyond one SGD update. After a single SGD update, the response in subsequent forward passes originates from the initial backward pass. Across multiple updates, however, each pass reuses weights that have already interacted with accumulated feature and gradient histories. Appendix F derives the corresponding finite-network and mean-field recursions. Passing formally from layer sums to depth integrals yields the candidate dynamics at training iterate k:

$$\mathrm{d}H_t^k = \sum_{s=0}^{k-1} \big[R_s^k(t) - \eta_c \chi_\infty^s C_\phi^{s,k}(t)\big] G_t^s \,\mathrm{d}t + \mathrm{d}W_t^k, \tag{21}$$

$$\mathrm{d}G_t^k = -\phi'(H_t^k) \left[\sum_{s=0}^{k} S_s^k(t)\phi(H_t^s) - \eta_c \sum_{s=0}^{k-1} \chi_\infty^s C_g^{s,k}(t)\phi(H_t^s)\right] \mathrm{d}t - \phi'(H_t^k)\,\mathrm{d}\widetilde{W}_t^k. \tag{22}$$

The coefficients R_s^k and S_s^k characterize the responses due to weight reuse, where the backward response incorporates the current feature $\phi(H_t^k)$ because the forward pass at iterate k has already been evaluated. The terms involving χ_∞^s (the limiting loss derivative at iterate s) capture direct SGD updates, while the cross-kernels $C_\phi^{s,k}(t) = \mathbb{E}[\phi(H_t^s)\phi(H_t^k)]$ and $C_g^{s,k}(t) = \mathbb{E}[G_t^s G_t^k]$ govern the covariance rates of the Gaussian drivers across training iterates. Boundary conditions and readout recursions are detailed in the appendix.

At fixed depth, width convergence requires tracking projections onto entire feature and gradient histories. When these histories are nearly linearly dependent, regression steps can amplify small errors in the empirical Gram matrices. Under the nondegeneracy condition in Eq. (197), Proposition F.4 recovers the $O(n^{-1/2})$ width rate for both passes at fixed L and k. Figure 3 examines this condition empirically by tracking the smallest eigenvalues of the history Gram matrices during training, showing that they remain positive early in training but degenerate toward zero as training progresses.

For depth convergence, the primary obstacle is the backward noise. Because the updated forward state H_t^k can depend on backward noise from earlier iterates at depths beyond t, and because the driving noise $\widetilde{W}^k$ shares those sources, $\phi'(H_t^k)$ may fail to be adapted to the filtration required for standard Brownian coupling. Consequently, the backward stochastic term remains formal for general nonlinear activations. For linear activations ($\phi(h) = h$), however, the Jacobian reduces to a constant, simplifying the backward noise to the increment $\widetilde{W}_1^k - \widetilde{W}_t^k$. Under a depth-uniform nondegeneracy condition on older histories, Theorem F.8 establishes an $O(L^{-1/2})$ depth rate for both passes, and Corollary F.10 combines the two estimates to yield the error bound $C_{p,L,k} n^{-1/2} + C_{p,k} L^{-1/2}$.

6 Experiments

Our experiments examine how approximation errors change with width and depth, how replacing the initial backward weights affects training and test losses, and how the feature and gradient histories become conditioned over the course of training. We primarily train ResNets of varying widths and depths on CIFAR-10 by SGD. Appendix G provides further experimental details.

Width and depth sweeps at initialization and after training. Figure 1 varies width and depth separately at initialization and after 30 epochs of SGD. At both checkpoints, the plotted approximation

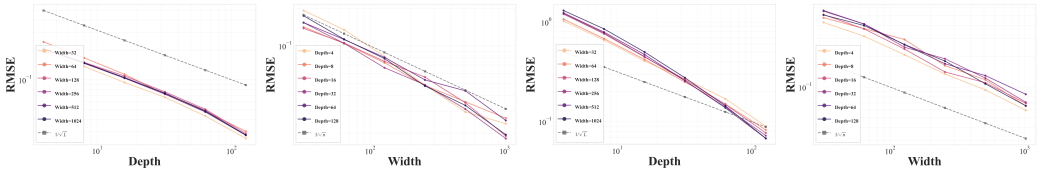
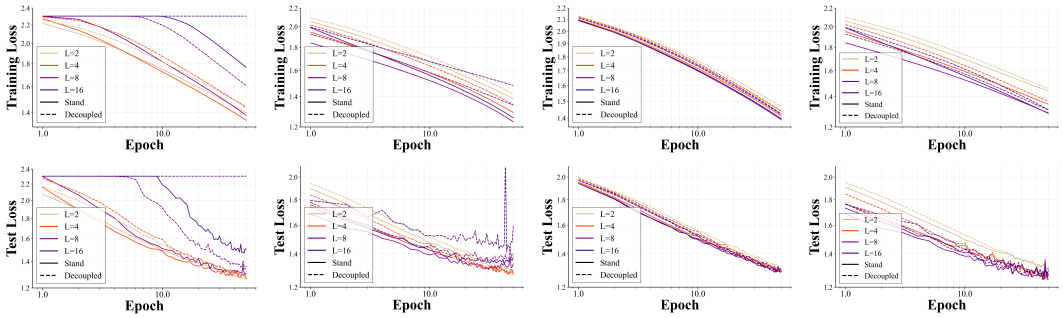


Figure 1: **Depth and width sweeps at initialization and after training.** The first two panels vary depth and width, respectively, at initialization. The last two repeat these sweeps after 30 epochs of SGD training on CIFAR-10.



(a) Vanilla DNN **(b) μP ResNet** **(c) Depth-μP ResNet** **(d) μP + weight decay**

Figure 2: **Loss curves with standard and decoupled backpropagation.** The top row shows training loss and the bottom row shows test loss. All networks have width 128. Colors indicate depth, and solid and dashed curves denote standard and decoupled backpropagation, respectively. In the decoupled runs, the backward weights start from an independent copy and receive the same subsequent updates as the forward weights in that run.

error is consistent with L^{-1} decay in the depth sweeps and n^{-1} decay in the width sweeps. At initialization, Theorem 4.3 gives L^p error bounds of $O(n^{-1/2} + L^{-1})$ for the forward process and $O(n^{-1/2} + L^{-1/2})$ for the backward process. After one SGD update, Theorem 5.3 gives an L^p bound of $C_{p,L}n^{-1/2} + C_pL^{-1/2}$ for the updated forward features and network output. The 30-epoch sweeps complement this one-update result with empirical rates over a longer training horizon.

Loss curves with decoupled backpropagation. Figure 2 compares standard and decoupled backpropagation across four settings. The loss curves track each other closely for depth-μP ResNets in panel (c). Panels (b) and (d) show μP ResNets without and with weight decay, respectively. In panel (d), the decoupled runs generally have slightly higher training loss, while the test-loss gaps narrow later in training. Our one-update analysis shows that the response induced by weight reuse can remain nonzero in the continuous-depth limit. The close loss curves in panels (c) and (d) suggest two possible explanations. The response may be small relative to the direct SGD drift, or it may change the features in ways that have little effect on the loss. Comparing panels (b) and (d) provides a starting point for studying how weight decay affects the response and its influence on the loss.

Conditioning of the feature and gradient histories. Figure 3 tracks the smallest eigenvalues of the empirical feature and gradient history matrices. The eigenvalues are positive early in training and approach zero as training progresses. Increasing width slows the decline over the range tested, while the downward trend persists. This deterioration helps explain why a uniform positive lower bound on the history eigenvalues can be restrictive for convergence analyses that cover multiple SGD updates.

7 Conclusions

In this work, we investigated infinite-depth feature learning in ResNets by isolating the forward–backward coupling induced by weight reuse. Our conditional Gaussian representations explicitly characterize this coupling at finite width and depth. At initialization, its effect on the backward process remains $O(n^{-1/2})$ in coordinatewise L^p uniformly across depth. Furthermore, for every

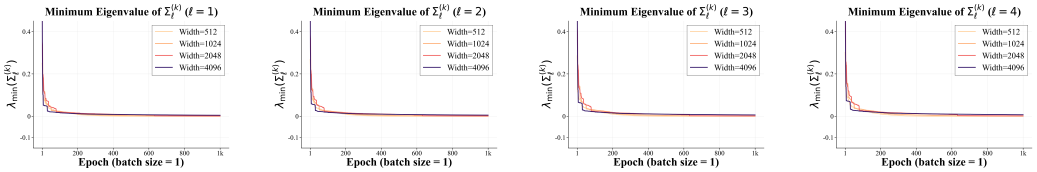


Figure 3: **Minimum empirical covariance eigenvalues during training.** The panels track the smallest eigenvalues of the empirical feature and gradient history covariance matrices by layer. ResNets have four hidden layers and widths from 512 to 4096. The runs use online SGD on CIFAR-10 across five seeds.

$p \geq 2$, the forward and backward processes converge jointly to their continuous-depth limits at coordinatewise L^p rates of $O(n^{-1/2} + L^{-1})$ and $O(n^{-1/2} + L^{-1/2})$, respectively.

Following a single SGD update, weight reuse introduces a response term in subsequent forward passes that persists in the depth limit at the same order as the direct SGD drift. We established the corresponding limiting dynamics and bounded the L^p approximation error for the updated features and network output by $C_{p,L}n^{-1/2} + C_pL^{-1/2}$. However, a linear ResNet counterexample demonstrates that updated feature moments can diverge with depth at fixed width, proving that the width and depth limits of these moments do not commute in general.

For arbitrary finite sequences of SGD updates, we derived mean-field recursions tracking accumulated feature and gradient histories. At fixed depth, nondegeneracy of the history Gram matrices yields an $O(n^{-1/2})$ width convergence rate, though empirical inspection shows these matrices degrade toward zero as training progresses. For linear activations, a depth-uniform lower eigenvalue bound on older histories establishes an $O(L^{-1/2})$ depth convergence rate for both passes. Combining these nondegeneracy conditions delivers the final width-first error bound of $C_{p,L,k}n^{-1/2} + C_{p,k}L^{-1/2}$.

Limitations. Our analysis focuses on ResNets with single-layer residual blocks trained by stream SGD under depth-μP. Over multiple updates, the width estimates require the accumulated feature and gradient histories to stay sufficiently far from linear dependence. The depth result covers linear activations and requires a lower eigenvalue bound on the older histories that is uniform in depth. For general nonlinear activations, the multistep continuous-depth equations remain candidates. How weight-reuse coupling affects optimization, learned representations, and generalization remains open.

Future directions. Two next steps are to extend the width estimates to poorly conditioned or singular histories and to justify the nonlinear backward stochastic term over multiple updates. Extending the conditional Gaussian approach to Transformers and adaptive optimizers could clarify when weight reuse persists across architectures and training rules, and how it affects optimization, learned representations, and generalization.

Acknowledgments and Disclosure of Funding

T.G. acknowledges support from DePaul University through a University Research Council Competitive Research Grant (Award No. 9710). Part of this research was performed while T.G. was visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation under Grant No. DMS-2425650. R.W. was partially supported by NSF DMS-2308120 and Simons Foundation travel grant MPTSM-00002346.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International conference on machine learning*, pages 322–332. PMLR, 2019.

[4] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[7] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*. PMLR, 2019.

[10] Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

[11] Tianxiang Gao. *Mastering infinite depths: Optimization and generalization in deeper neural networks*. PhD thesis, Iowa State University, 2024.

[12] Tianxiang Gao, Xiaokai Huo, Hailiang Liu, and Hongyang Gao. Wide neural networks as gaussian processes: Lessons from deep equilibrium models. *Advances in Neural Information Processing Systems*, 2023.

[13] Tianxiang Gao, Hailiang Liu, Jia Liu, Hridesh Rajan, and Hongyang Gao. A global convergence theory for deep relu implicit networks via over-parameterization. In *International Conference on Learning Representations*, 2021.

[14] Tianxiang Gao, Siyuan Sun, Hailiang Liu, and Hongyang Gao. Global convergence in neural odes: Impact of activation functions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[15] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning*. PMLR, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016.

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

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[20] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[21] 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*, 2022.

[22] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[23] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *Mathematical Statistics and Learning*, 6(3):201–357, 2023.

[24] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

[25] Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

[26] Huy Tuan Pham and Phan-Minh Nguyen. Global convergence of three-layer neural networks in the mean field regime. In *International Conference on Learning Representations*, 2021.

[27] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 2016.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[29] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

[30] Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, et al. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026.

[31] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

[34] Greg Yang. Tensor programs I: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 2019.

[35] Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

[36] Greg Yang. Tensor programs III: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

[37] Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[38] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, 2021.

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

[40] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 2020.

A Useful Mathematical Results

Lemma A.1 (Gronwall's inequality)**.** *Let $I = [a, b]$ for an interval such that $a < b < \infty$. Let u, α, β be real-valued continuous functions such that β is non-negative and u satisfies the integral inequality*

$$u(t) \leq \alpha(t) + \int_0^t \beta(s)u(s)ds, \quad \forall t \in I.$$

Then

$$u(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp\left(\int_s^t \beta(r)dr\right), \quad \forall t \in I.$$

If, in addition, $\alpha(t)$ is non-decreasing, then

$$u(t) \leq \alpha(t) \exp\left(\int_0^t \beta(s)ds\right), \quad \forall t \in I.$$

Lemma A.2 (Gronwall's inequality (discrete version))**.** *Let (u_n) and (β_n) be non-negative sequences satisfying*

$$u_n \leq \alpha + \sum_{k=0}^{n-1} \beta_k u_k, \quad \forall\, n,$$

where $\alpha \geq 0$. Then

$$u_n \leq \alpha \exp\left(\sum_{k=0}^{n-1} \beta_k\right), \quad \forall\, n.$$

Lemma A.3 (Gaussian increment estimate)**.** *Let $Y \in \mathbb{R}^d$ be deterministic and let $Z \in \mathbb{R}^d$ be centered Gaussian. Set $s^2 := \mathbb{E}\|Z\|^2$. For every $p \geq 2$, there is a constant C_p, depending only on p, such that, for every deterministic $\varepsilon \in [0, 1]$,*

$$\mathbb{E}\|Y + \varepsilon Z\|^p \leq (1 + C_p\varepsilon^2)\|Y\|^p + C_p\varepsilon^2 s^p. \tag{23}$$

The constant C_p is independent of d and the covariance of Z.

Proof. For $p = 2$, expand the square and use $\mathbb{E}Z = 0$. For $p > 2$, the Taylor bound give

$$\mathbb{E}\|Y + \varepsilon Z\|^p \leq \|Y\|^p + C_p\varepsilon^2\|Y\|^{p-2}s^2 + C_p\varepsilon^p s^p,$$

where we used the Gaussian moment bound $\mathbb{E}\|Z\|^p \leq C_p s^p$. Young's inequality bounds $\|Y\|^{p-2}s^2$ by $C_p(\|Y\|^p + s^p)$. Since $\varepsilon^p \leq \varepsilon^2$, this proves Eq. (23). □

Lemma A.4 (Rosenthal's inequality)**.** *Let $\{X_j\}_{j=1}^m$ be independent centered real random variables with finite q-th moments, where $q \geq 2$. There is a constant C_q, depending only on q, such that*

$$\mathbb{E}\left|\sum_{j=1}^m X_j\right|^q \leq C_q \max\left\{\sum_{j=1}^m \mathbb{E}|X_j|^q,\ \left(\sum_{j=1}^m \mathbb{E}X_j^2\right)^{q/2}\right\}. \tag{24}$$

Lemma A.5 (Empirical product estimate)**.** *Let $p \geq 2$ and let $(\bar{X}_i, \bar{Y}_i)_{i=1}^n$ be iid pairs with uniformly bounded L^{2p} moments. Suppose their coupled variables satisfy*

$$\sup_i\big(\|X_i - \bar{X}_i\|_{L^{2p}} + \|Y_i - \bar{Y}_i\|_{L^{2p}}\big) \leq Cn^{-1/2}.$$

Then

$$\left\|\frac{1}{n}\sum_{i=1}^n X_iY_i - \mathbb{E}[\bar{X}_1\bar{Y}_1]\right\|_{L^p} \leq C_p n^{-1/2}, \tag{25}$$

where C_p depends only on p and the assumed bounds.

Proof. The identity

$$X_iY_i - \bar{X}_i\bar{Y}_i = (X_i - \bar{X}_i)Y_i + \bar{X}_i(Y_i - \bar{Y}_i)$$

and Hölder's inequality give $\|X_iY_i - \bar{X}_i\bar{Y}_i\|_{L^p} \le C_p n^{-1/2}$. Minkowski's inequality gives the same bound for their average. The centered products $\bar{X}_i\bar{Y}_i - \mathbb{E}[\bar{X}_i\bar{Y}_i]$ are iid with uniformly bounded p-th moments, so Rosenthal's inequality from Lemma A.4 bounds their average by $C_p n^{-1/2}$ in L^p. Adding the two bounds proves the claim. □

Lemma A.6 (Gaussian martingale moment bound)**.** *Let $(\mathcal{F}_j)_{j=0}^m$ be a filtration, and let X_j be $\mathcal{F}_j$-measurable and conditionally centered Gaussian given $\mathcal{F}_{j-1}$. Set*

$$s_j^2 := \mathbb{E}[X_j^2 \mid \mathcal{F}_{j-1}].$$

For every $q \ge 2$, there is a constant C_q, depending only on q, such that

$$\mathbb{E}\left|\sum_{j=1}^m X_j\right|^q \le C_q \mathbb{E}\left(\sum_{j=1}^m s_j^2\right)^{q/2}, \tag{26}$$

provided the right-hand side is finite.

Lemma A.7 (Picard–Lindelöf theorem)**.** *Let $U \subseteq \mathbb{R}^d$ be open and $T \in (0,\infty]$. Let $f : [0,T) \times U \to \mathbb{R}^d$ be continuous and locally Lipschitz in x, locally uniformly in t. For every $x_0 \in U$, the problem*

$$x'(t) = f(t, x(t)), \qquad x(0) = x_0,$$

has a unique maximal solution $x \in C^1([0,\tau); U)$ with $\tau \in (0,T]$. If $\tau < T$, then $x(t)$ eventually lies outside each compact subset of U as $t \uparrow \tau$.

In particular, if $U = \mathbb{R}^d$ and, for every $S < T$, there is a constant C_S such that

$$\|f(t,x)\| \le C_S(1 + \|x\|), \qquad (t,x) \in [0,S] \times \mathbb{R}^d,$$

then $\tau = T$.

Lemma A.8 (Euler local truncation estimates)**.** *Let $U \subseteq \mathbb{R}^d$ be open, let $f : [0,T] \times U \to \mathbb{R}^d$ be locally Lipschitz in time and state, and suppose that $x \in C^1([0,T]; U)$ solves*

$$x'(t) = f(t, x(t)).$$

For the uniform grid $t_k = k\Delta$, where $\Delta = T/N$,

$$x(t_k) = x(t_{k-1}) + \Delta f(t_{k-1}, x(t_{k-1})) + R_k,$$
$$x(t_{k-1}) = x(t_k) - \Delta f(t_{k-1}, x(t_k)) + \widetilde{R}_k,$$

where

$$\max_{1 \le k \le N} \big(\|R_k\| + \|\widetilde{R}_k\|\big) \le C\Delta^2.$$

The constant C is independent of N.

Lemma A.9 (Burkholder–Davis–Gundy inequality)**.** *Let B be a standard Brownian motion, let $p \ge 2$, and let a be a predictable process satisfying*

$$\mathbb{E}\left(\int_0^T |a_s|^2\, ds\right)^{p/2} < \infty.$$

There is a constant C_p, depending only on p, such that

$$\mathbb{E}\left[\sup_{0 \le t \le T}\left|\int_0^t a_s\, dB_s\right|^p\right] \le C_p \mathbb{E}\left(\int_0^T |a_s|^2\, ds\right)^{p/2}.$$

In particular, for $0 \le s \le t \le T$,

$$\mathbb{E}\left|\int_s^t a_u\, dB_u\right|^p \le C_p \mathbb{E}\left(\int_s^t |a_u|^2\, du\right)^{p/2}.$$

B Post-Activation vs. Pre-Activation ResNet Design

The ResNet architecture defined in Eq. (1) follows the so-called *pre-activation style* [17], where the residual connection is applied to the *pre-activation* feature vectors. There is, however, another widely used variant: the *post-activation style* [16], in which the residual connection is applied after the nonlinearity. The corresponding residual recursion is

$$h_0 = \phi\left(\frac{1}{\sqrt{d}}Ux\right), \quad h_\ell = h_{\ell-1} + \frac{1}{\sqrt{L}}\,\phi\left(\frac{1}{\sqrt{n}}W_\ell h_{\ell-1}\right), \quad \forall \ell \in \{1, 2, \dots, L\}. \tag{27}$$

Historically, the post-activation formulation was first introduced in the original ResNet paper [16], but subsequent large-scale architectures, including modern Transformers [32], have overwhelmingly favored the pre-activation design. This shift was motivated primarily by the empirically observed superior stability of pre-activation models. However, the theoretical reasons for this preference are less often articulated. Proposition 3.1 shows that pre-activation is essentially more stable as the depth $L \to \infty$ by proving that the feature magnitudes in the post-activation variant Eq. (27) may diverge, even under the $\frac{1}{\sqrt{L}}$ depth-scaling.

Proof of Proposition 3.1. Recall that ϕ is assumed to be positive dominate, that is, there exist non-negative constants c_1, c_2, not both zero, such that $\mathbb{E}[\phi(xZ)] \geq c_1|x| + c_2$, $\forall x \in \mathbb{R}$, where Z is the standard Gaussian random variable.

First note that $(Ux)_i \sim \mathcal{N}(0, \|x\|^2)$ and hence $\mathbb{E}[h_{0,i}] \geq c_1 \frac{\|x\|}{\sqrt{d}} + c_2$. Let $\mathcal{B}_\ell$ be the σ-algebra generated by $\{h_0, \cdots, h_{\ell-1}\}$. Then, observe that

$$\begin{aligned}
\mathbb{E}[h_{\ell,i} \mid \mathcal{B}_\ell] &= h_{\ell-1,i} + \frac{1}{\sqrt{L}}\mathbb{E}\left[\phi\left(\frac{\|h_{\ell-1}\|}{\sqrt{n}}Z\right) \mid \mathcal{B}_\ell\right] \geq h_{\ell-1,i} + \frac{1}{\sqrt{L}}\left(c_1\frac{\|h_{\ell-1}\|}{\sqrt{n}} + c_2\right) \\
&\geq h_{\ell-1,i} + \frac{1}{\sqrt{L}}\left(c_1\frac{h_{\ell-1,i}}{\sqrt{n}} + c_2\right) = \left(1 + c_1\frac{1}{\sqrt{Ln}}\right)h_{\ell-1,i} + c_2\frac{1}{\sqrt{L}}.
\end{aligned}$$

Then taking expectation of $\mathcal{B}_\ell$ yields

$$\mathbb{E}[h_{\ell,i}] \geq \left(1 + c_1\frac{1}{\sqrt{Ln}}\right)\mathbb{E}[h_{\ell-1,i}] + c_2\frac{1}{\sqrt{L}}.$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{E}[h_{\ell,i}] &\geq \left(1 + c_1\frac{1}{\sqrt{Ln}}\right)^\ell \mathbb{E}[h_{0,i}] + c_2\frac{1}{\sqrt{L}}\sum_{j=0}^{\ell-1}\left(1 + c_1\frac{1}{\sqrt{Ln}}\right)^j \\
&\geq \left(1 + c_1\frac{1}{\sqrt{Ln}}\right)^\ell \left(c_1\frac{\|x\|}{\sqrt{d}} + c_2\right) + c_2\frac{1}{\sqrt{L}}\ell.
\end{aligned}$$

Therefore we obtain

$$\mathbb{E}[h_{L,i}] \geq c_1\frac{\|x\|}{\sqrt{d}}\left(1 + c_1\frac{1}{\sqrt{Ln}}\right)^L + c_2\sqrt{TL} \to \infty$$

as $L \to \infty$, provided either $c_1 > 0$ or $c_2 > 0$. □

C Gaussian Representation of Forward and Backward Innovations

We derive conditional Gaussian representations of the initialized-weight queries after an arbitrary number of SGD updates. The formulas separate the contributions determined by previous queries from the remaining Gaussian innovations and will be used directly in the width-convergence proofs.

At training iterate $j \geq 0$, the hidden weights satisfy

$$W_\ell^j = W_\ell^0 - \frac{\eta_c}{\sqrt{Ln}}\sum_{s<j}\chi^s g_\ell^s(\phi_{\ell-1}^s)^\top. \tag{28}$$

Define the initialized-weight queries

$$a_\ell^j := \frac{1}{\sqrt{n}} W_\ell^0 \phi_{\ell-1}^j, \qquad b_\ell^j := \frac{1}{\sqrt{n}} (W_\ell^0)^\top g_\ell^j, \tag{29}$$

where $\phi_{\ell-1}^j := \phi(h_{\ell-1}^j)$.

Collect the previous query directions and evaluations into

$$\Phi_{\ell-1}^{<j} := [\phi_{\ell-1}^0 \ \cdots \ \phi_{\ell-1}^{j-1}], \qquad A_\ell^{<j} := [a_\ell^0 \ \cdots \ a_\ell^{j-1}], \tag{30}$$

$$G_\ell^{<j} := [g_\ell^0 \ \cdots \ g_\ell^{j-1}], \qquad B_\ell^{<j} := [b_\ell^0 \ \cdots \ b_\ell^{j-1}], \tag{31}$$

all in $\mathbb{R}^{n\times j}$. For ordered sets of training indices I, J, define

$$C_{\phi,\ell-1}^{I,J} := \frac{1}{n} (\Phi_{\ell-1}^I)^\top \Phi_{\ell-1}^J, \qquad C_{g,\ell}^{I,J} := \frac{1}{n} (G_\ell^I)^\top G_\ell^J.$$

Thus, $C_{\theta,\ell}^{I,J} \in \mathbb{R}^{|I|\times|J|}$, with a single index interpreted as a singleton.

Define the residual query directions

$$\begin{aligned} \phi_{\ell-1,\perp}^j &:= (I - P_{\Phi_{\ell-1}^{<j}}) \phi_{\ell-1}^j \\ &= \phi_{\ell-1}^j - \Phi_{\ell-1}^{<j} (C_{\phi,\ell-1}^{<j,<j})^\dagger C_{\phi,\ell-1}^{<j,j}, \end{aligned} \tag{32}$$

$$\begin{aligned} g_{\ell,\perp}^j &:= (I - P_{G_\ell^{<j}}) g_\ell^j \\ &= g_\ell^j - G_\ell^{<j} (C_{g,\ell}^{<j,<j})^\dagger C_{g,\ell}^{<j,j}. \end{aligned} \tag{33}$$

where $M^\dagger$ is Moore–Penrose pseudoinverse of a matrix M, and $P_M := MM^\dagger$ the orthogonal projection onto its column span. Empty histories have zero projections, and their matrix-product contributions are interpreted as zero.

Substituting Eq. (28) into the network gives

$$h_\ell^j = h_{\ell-1}^j + \frac{1}{\sqrt{L}} a_\ell^j - \frac{\eta_c}{L} \sum_{s<j} \chi^s C_{\phi,\ell-1}^{s,j} g_\ell^s, \tag{34}$$

$$g_{\ell-1}^j = g_\ell^j + \left[\frac{1}{\sqrt{L}} b_\ell^j - \frac{\eta_c}{L} \sum_{s<j} \chi^s C_{g,\ell}^{s,j} \phi_{\ell-1}^s \right] \odot \phi'(h_{\ell-1}^j). \tag{35}$$

The computation history consists of the initial input and readout weights and the query directions and evaluations already revealed. At each training iterate, forward queries are evaluated in increasing layer order and backward queries in decreasing layer order. In particular, before evaluating b_ℓ^j, the current forward query a_ℓ^j has already been revealed.

Lemma C.1 (Conditional Gaussian queries of the initialized weights)**.** *For every $\ell \in [L]$ and $j \geq 0$, the following conditional representations hold.*

(i) ***Forward query.*** *Conditionally on the computation history before evaluating a_ℓ^j,*

$$\begin{aligned} a_\ell^j \overset{d}{=} & A_\ell^{<j} (C_{\phi,\ell-1}^{<j,<j})^\dagger C_{\phi,\ell-1}^{<j,j} \\ & + G_\ell^{<j} (C_{g,\ell}^{<j,<j})^\dagger \frac{1}{n} (B_\ell^{<j})^\top \phi_{\ell-1,\perp}^j \\ & + \frac{\|\phi_{\ell-1,\perp}^j\|}{\sqrt{n}} (I - P_{G_\ell^{<j}}) z_\ell^j, \end{aligned} \tag{36}$$

where $z_\ell^j \sim \mathcal{N}(0, I_n)$ is independent of that history.

(ii) ***Backward query.*** *Conditionally on the computation history before evaluating b_ℓ^j,*

$$\begin{aligned} b_\ell^j \overset{d}{=} & B_\ell^{<j} (C_{g,\ell}^{<j,<j})^\dagger C_{g,\ell}^{<j,j} \\ & + \Phi_{\ell-1}^{\leq j} (C_{\phi,\ell-1}^{\leq j,\leq j})^\dagger \frac{1}{n} (A_\ell^{\leq j})^\top g_{\ell,\perp}^j \\ & + \frac{\|g_{\ell,\perp}^j\|}{\sqrt{n}} (I - P_{\Phi_{\ell-1}^{\leq j}}) \tilde{z}_\ell^j, \end{aligned} \tag{37}$$

where $\widetilde{z}_\ell^j \sim \mathcal{N}(0, I_n)$ is independent of that history.

These conditional laws admit a joint realization with mutually independent standard Gaussian innovations $\{z_\ell^j, \widetilde{z}_\ell^j\}_{\ell,j}$ introduced in computation order. Whenever the indicated empirical Gram matrices are invertible, their pseudoinverses in Eq. (36) and Eq. (37) are ordinary inverses.

Proof. Before evaluating a_ℓ^j, the recorded queries impose

$$W_\ell^0 \Phi_{\ell-1}^{<j} = \sqrt{n}\, A_\ell^{<j}, \qquad (W_\ell^0)^\top G_\ell^{<j} = \sqrt{n}\, B_\ell^{<j}. \tag{38}$$

The initialized matrix has the exact decomposition

$$W_\ell^0 = W_\ell^0 P_{\Phi_{\ell-1}^{<j}} + P_{G_\ell^{<j}} W_\ell^0 (I - P_{\Phi_{\ell-1}^{<j}}) + (I - P_{G_\ell^{<j}}) W_\ell^0 (I - P_{\Phi_{\ell-1}^{<j}}). \tag{39}$$

The first two terms are determined by the recorded constraints:

$$W_\ell^0 P_{\Phi_{\ell-1}^{<j}} = \sqrt{n}\, A_\ell^{<j} (\Phi_{\ell-1}^{<j})^\dagger, \quad P_{G_\ell^{<j}} W_\ell^0 (I - P_{\Phi_{\ell-1}^{<j}}) = \sqrt{n}\, (G_\ell^{<j})^{\dagger\top} (B_\ell^{<j})^\top (I - P_{\Phi_{\ell-1}^{<j}}).$$

As each query direction is determined before its evaluation, successive Gaussian conditioning gives

$$W_\ell^0 \overset{d}{=} \sqrt{n}\, A_\ell^{<j} (\Phi_{\ell-1}^{<j})^\dagger + \sqrt{n}\, (G_\ell^{<j})^{\dagger\top} (B_\ell^{<j})^\top (I - P_{\Phi_{\ell-1}^{<j}}) + (I - P_{G_\ell^{<j}}) \widetilde{W}_\ell^j (I - P_{\Phi_{\ell-1}^{<j}}), \tag{40}$$

conditionally on the computation history, where $\widetilde{W}_\ell^j$ has iid standard Gaussian entries and is independent of that history.

Multiplying Eq. (40) by $\phi_{\ell-1}^j / \sqrt{n}$ gives the conditional mean

$$A_\ell^{<j} (\Phi_{\ell-1}^{<j})^\dagger \phi_{\ell-1}^j + (G_\ell^{<j})^{\dagger\top} (B_\ell^{<j})^\top \phi_{\ell-1,\perp}^j.$$

The pseudoinverse identities

$$(\Phi_{\ell-1}^{<j})^\dagger = \frac{1}{n} (C_{\phi,\ell-1}^{<j,<j})^\dagger (\Phi_{\ell-1}^{<j})^\top, \qquad (G_\ell^{<j})^{\dagger\top} = \frac{1}{n} G_\ell^{<j} (C_{g,\ell}^{<j,<j})^\dagger$$

give the first two terms of Eq. (36). The remaining term is conditionally centered Gaussian with covariance

$$\frac{\|\phi_{\ell-1,\perp}^j\|^2}{n} (I - P_{G_\ell^{<j}}),$$

which gives its stated Gaussian representation.

Before evaluating b_ℓ^j, the current forward query has also revealed

$$W_\ell^0 \phi_{\ell-1}^j = \sqrt{n}\, a_\ell^j.$$

The forward history is therefore $(\Phi_{\ell-1}^{\leq j}, A_\ell^{\leq j})$, while the backward history remains $(G_\ell^{<j}, B_\ell^{<j})$. Applying the same decomposition to $(W_\ell^0)^\top$ gives the conditional mean

$$B_\ell^{<j} (G_\ell^{<j})^\dagger g_\ell^j + (\Phi_{\ell-1}^{\leq j})^{\dagger\top} (A_\ell^{\leq j})^\top g_{\ell,\perp}^j.$$

Its residual covariance is

$$\frac{\|g_{\ell,\perp}^j\|^2}{n} (I - P_{\Phi_{\ell-1}^{\leq j}}).$$

The same pseudoinverse identities yield Eq. (37).

Finally, sampling each conditional query with a fresh standard Gaussian vector in computation order preserves the joint law and gives the asserted independent innovation family. □

Proposition C.2 (Gaussian representation during training)**.** *The training recursions Eq. (34) and Eq. (35) admit a joint realization in which the initialized-weight queries a_ℓ^j, b_ℓ^j are given by Eq. (36) and Eq. (37), respectively. Each Gaussian innovation is independent of the computation history before its query.*

Proof. Substitute the conditional query representations into the state recursions successively in computation order. All other terms are determined before the corresponding query. The joint realization follows from Lemma C.1. □

Corollary C.3 (Gaussian representation at initialization)**.** *At initialization, the conditional query representations reduce to*

$$a_\ell^0 \overset{d}{=} \frac{\|\phi_{\ell-1}^0\|}{\sqrt{n}} z_\ell^0, \tag{41}$$

$$b_\ell^0 \overset{d}{=} \frac{\langle a_\ell^0, g_\ell^0 \rangle}{\|\phi_{\ell-1}^0\|^2} \phi_{\ell-1}^0 + \frac{\|g_\ell^0\|}{\sqrt{n}} (I - P_{\phi_{\ell-1}^0}) \widetilde{z}_\ell^0. \tag{42}$$

When $\phi_{\ell-1}^0 = 0$, its projection and the aligned term are interpreted as zero. In the joint realization,

$$h_\ell^0 = h_{\ell-1}^0 + \frac{1}{\sqrt{L}} a_\ell^0, \tag{43}$$

$$g_{\ell-1}^0 = g_\ell^0 + \frac{1}{\sqrt{L}} b_\ell^0 \odot \phi'(h_{\ell-1}^0). \tag{44}$$

Proof. Set $j = 0$ in Lemma C.1. The previous histories are empty, while the backward query uses $\Phi_{\ell-1}^{\leq 0} = [\phi_{\ell-1}^0]$ and $A_\ell^{\leq 0} = [a_\ell^0]$. The state recursions have empty SGD correction sums. □

Corollary C.4 (Gaussian representation after one SGD update)**.** *After one update, the conditional forward query representation is*

$$a_\ell^1 \overset{d}{=} \frac{C_{\phi,\ell-1}^{0,1}}{C_{\phi,\ell-1}^{0,0}} a_\ell^0 + \frac{\langle b_\ell^0, \phi_{\ell-1,\perp}^1 \rangle}{\|g_\ell^0\|^2} g_\ell^0 + \frac{\|\phi_{\ell-1,\perp}^1\|}{\sqrt{n}} (I - P_{g_\ell^0}) z_\ell^1, \tag{45}$$

and the conditional backward query representation is

$$b_\ell^1 \overset{d}{=} \frac{C_{g,\ell}^{0,1}}{C_{g,\ell}^{0,0}} b_\ell^0 + \Phi_{\ell-1}^{\leq 1} (C_{\phi,\ell-1}^{\leq 1,\leq 1})^\dagger \frac{1}{n} (A_\ell^{\leq 1})^\top g_{\ell,\perp}^1 + \frac{\|g_{\ell,\perp}^1\|}{\sqrt{n}} (I - P_{\Phi_{\ell-1}^{\leq 1}}) \widetilde{z}_\ell^1. \tag{46}$$

Here

$$\phi_{\ell-1,\perp}^1 = (I - P_{\phi_{\ell-1}^0}) \phi_{\ell-1}^1, \qquad g_{\ell,\perp}^1 = (I - P_{g_\ell^0}) g_\ell^1.$$

Ratios with zero denominators are interpreted as zero. The corresponding state recursions are

$$h_\ell^1 = h_{\ell-1}^1 + \frac{1}{\sqrt{L}} a_\ell^1 - \frac{\eta_c}{L} \chi^0 C_{\phi,\ell-1}^{0,1} g_\ell^0, \tag{47}$$

$$g_{\ell-1}^1 = g_\ell^1 + \left[\frac{1}{\sqrt{L}} b_\ell^1 - \frac{\eta_c}{L} \chi^0 C_{g,\ell}^{0,1} \phi_{\ell-1}^0 \right] \odot \phi'(h_{\ell-1}^1). \tag{48}$$

Proof. Set $j = 1$ in Lemma C.1 and in the state recursions. Each history with superscript < 1 has one column, so its Gram pseudoinverse reduces to the reciprocal of its scalar Gram entry, with the zero convention. □

D Convergence Proofs at Initialization

We first establish the width limit of the forward and backward processes, uniformly over depth. The backward argument compares the original process with an auxiliary process driven by independent backward weights. We then establish well-posedness of the limiting forward–backward system and quantify the depth approximation. The width estimates have order $n^{-1/2}$ in L^p, while the forward and backward depth estimates have orders L^{-1} and $L^{-1/2}$, respectively.

Throughout this section, the training input $x^0 \in \mathbb{R}^d$ is deterministic and $q_0 := \|x^0\|^2/d > 0$. We assume that the nonlinear ϕ is globally Lipschitz. For the backward width and depth limits, we additionally assume that $\phi \in C^1(\mathbb{R})$ and ϕ' is globally Lipschitz. Write K_0 for a Lipschitz constant

of ϕ and K_1 for a Lipschitz constant of ϕ'. Constants C, C_p may depend on q_0, ϕ and the moment order, but not on n or L. We write $\|Y\|_{L^p} := (\mathbb{E}\|Y\|^p)^{1/p}$, using the Euclidean norm for vectors.

Across our analysis, the superscript k denotes the parameter state after the k-th update. Finite-width states are written as $h_{\ell,i}^{k,n,L}$, discrete mean fields as $\bar{h}_{\ell,i}^{k,L}$, and continuous-depth limits as $H_{t,i}^{k}$. For activated features, write

$$\phi_{\ell,i}^{k,n,L} := \phi(h_{\ell,i}^{k,n,L}), \qquad \bar{\phi}_{\ell,i}^{k,L} := \phi(\bar{h}_{\ell,i}^{k,L}).$$

D.1 Discrete Mean Fields and Moment Bounds

At initialization, the finite-width states are denoted by $h_{\ell,i}^{0,n,L}$ and $g_{\ell,i}^{0,n,L}$, and their discrete-depth mean fields by $\bar{h}_{\ell,i}^{0,L}$ and $\bar{g}_{\ell,i}^{0,L}$. For each coordinate i, the mean-field recursions are

$$\bar{h}_{0,i}^{0,L} = h_{0,i}^{0,n,L} \sim \mathcal{N}(0, q_0), \qquad \bar{h}_{\ell,i}^{0,L} = \bar{h}_{\ell-1,i}^{0,L} + \frac{1}{\sqrt{L}}\xi_{\ell,i}^{0,L}, \tag{49}$$

$$\bar{g}_{L,i}^{0,L} = v_i^0 \sim \mathcal{N}(0,1), \qquad \bar{g}_{\ell-1,i}^{0,L} = \bar{g}_{\ell,i}^{0,L} + \frac{1}{\sqrt{L}}\phi'(\bar{h}_{\ell-1,i}^{0,L})\zeta_{\ell,i}^{0,L}, \tag{50}$$

where

$$\xi_{\ell,i}^{0,L} := \sqrt{\bar{C}_{\phi,\ell-1}^{0,0,L}}\, z_{\ell,i}^0, \qquad \bar{C}_{\phi,\ell}^{0,0,L} := \mathbb{E}[(\bar{\phi}_{\ell,i}^{0,L})^2], \tag{51}$$

$$\zeta_{\ell,i}^{0,L} := \sqrt{\bar{C}_{g,\ell}^{0,0,L}}\, \widetilde{z}_{\ell,i}^0, \qquad \bar{C}_{g,\ell}^{0,0,L} := \mathbb{E}[(\bar{g}_{\ell,i}^{0,L})^2]. \tag{52}$$

The standard Gaussian variables $z_{\ell,i}^0, \widetilde{z}_{\ell,i}^0$ are mutually independent across layers, coordinates, and families, and independent of the initial states and v^0. Thus, the joint pairs $(\bar{h}_{\cdot,i}^{0,L}, \bar{g}_{\cdot,i}^{0,L})$ are independent copies across i. The forward and backward processes within a pair need not be independent.

Lemma D.1 (Uniform moments and nondegeneracy)**.** *Suppose that $q_0 > 0$ and that ϕ is globally K_0-Lipschitz and not identically zero. For the backward assertions, assume additionally that ϕ is differentiable; then $\|\phi'\|_\infty \le K_0$. There are constants $c, C, Q > 0$ such that, uniformly in L and ℓ,*

$$q_0 \le \mathbb{E}[(\bar{h}_{\ell,i}^{0,L})^2] \le Q, \qquad c \le \bar{C}_{\phi,\ell}^{0,0,L} \le C, \qquad 1 \le \bar{C}_{g,\ell}^{0,0,L} \le e^{K_0^2}. \tag{53}$$

For every $p \ge 2$,

$$\sup_{L\ge 1} \sup_{0\le \ell \le L} \sup_i \left(\|\bar{h}_{\ell,i}^{0,L}\|_{L^p} + \|\bar{\phi}_{\ell,i}^{0,L}\|_{L^p} + \|\bar{g}_{\ell,i}^{0,L}\|_{L^p} \right) \le C_p. \tag{54}$$

Proof. The forward innovations $\xi_{\ell,i}^{0,L}$ are centered Gaussian with deterministic variances, so $\bar{h}_{\ell,i}^{0,L}$ is centered Gaussian as well and

$$\mathbb{E}[(\bar{h}_{\ell,i}^{0,L})^2] = \mathbb{E}[(\bar{h}_{\ell-1,i}^{0,L})^2] + \frac{1}{L}\mathbb{E}[(\bar{\phi}_{\ell-1,i}^{0,L})^2].$$

The variance is therefore at least q_0. Lipschitz continuity of ϕ gives the bound $|\phi(u)| \le K_0|u| + |\phi(0)|$. Substituting this into the variance recursion yields

$$\mathbb{E}\big[(\bar{h}_{\ell,i}^{0,L})^2\big] \le \left(1 + \frac{2K_0^2}{L}\right) \mathbb{E}\big[(\bar{h}_{\ell-1,i}^{0,L})^2\big] + \frac{2|\phi(0)|^2}{L}.$$

Discrete Gronwall therefore gives a uniform upper bound.

We next bound the feature second moment from below. Since ϕ is continuous and not identically zero, it is bounded away from zero on some interval. Every nondegenerate Gaussian distribution assigns positive probability to this interval, so

$$\mathbb{E}[\phi(\sqrt{q}Z)^2] > 0, \qquad q > 0, \quad Z \sim \mathcal{N}(0,1).$$

The linear-growth bound on ϕ and dominated convergence show that this expectation is continuous in q. It therefore attains a strictly positive minimum on the compact range $[q_0, Q]$. Since $\bar{h}_{\ell,i}^{0,L}$ is centered Gaussian with variance in this $[q_0, Q]$, we obtain the uniform lower bound on $\bar{C}_{\phi,\ell}^{0,0,L}$.

Finally, for every $p \geq 2$, the Gaussian moment formula and uniform variance bound give an L^p bound on $\bar{h}^{0,L}_{\ell,i}$. Since ϕ has linear growth, this also gives the corresponding L^p bound on $\bar{\phi}^{0,L}_{\ell,i}$.

For the backward process, centering of the current innovation gives

$$\bar{C}^{0,0,L}_{g,\ell-1} = \left(1 + \frac{1}{L}\mathbb{E}[\phi'(\bar{h}^{0,L}_{\ell-1,i})^2]\right)\bar{C}^{0,0,L}_{g,\ell}.$$

Iterating from $\bar{C}^{0,0,L}_{g,L} = 1$ proves the variance bounds. Conditionally on the full mean-field forward path, $\bar{g}^{0,L}_{\ell,i}$ is centered Gaussian with variance

$$1 + \frac{1}{L}\sum_{j=\ell+1}^{L} \phi'(\bar{h}^{0,L}_{j-1,i})^2 \bar{C}^{0,0,L}_{g,j} \leq 1 + K_0^2 e^{K_0^2}.$$

The remaining moment bound follows from Gaussian moments. □

D.2 Width Convergence of the Forward Process

Proposition D.2 (Forward width convergence)**.** *Suppose that $q_0 > 0$ and that ϕ is globally Lipschitz and not identically zero. For every $p \geq 2$, there is a coupling such that*

$$\sup_{L\geq 1}\sup_{0\leq\ell\leq L}\sup_{1\leq i\leq n} \|h^{0,n,L}_{\ell,i} - \bar{h}^{0,L}_{\ell,i}\|_{L^p} \leq C_p n^{-1/2}.$$

The coupled finite-width features have uniformly bounded coordinate moments of every finite order.

Proof. We suppress n, L in the process notation, but retain the training-step superscripts. Couple the initial states by $h^0_{0,i} = \bar{h}^0_{0,i}$ and use the same forward standard Gaussian vectors in the finite and mean-field recursions:

$$a^0_\ell = \sqrt{C^{0,0}_{\phi,\ell-1}}\, z^0_\ell, \qquad \xi^0_\ell = \sqrt{\bar{C}^{0,0}_{\phi,\ell-1}}\, z^0_\ell.$$

At each layer, z^0_ℓ is independent of the preceding forward history. When $C^{0,0}_{\phi,\ell-1} = 0$, take z^0_ℓ to be an independent standard Gaussian vector and use it in both recursions. Define

$$e_\ell := \left(\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} |h^0_{\ell,i} - \bar{h}^0_{\ell,i}|^p\right)^{1/p}.$$

The reverse triangle inequality and the Lipschitz property give

$$\left|\sqrt{C^{0,0}_{\phi,\ell}} - \sqrt{\bar{C}^{0,0}_{\phi,\ell}}\right| \leq \frac{K_0}{\sqrt{n}}\|h^0_\ell - \bar{h}^0_\ell\| + \left|\frac{\|\bar{\phi}^0_\ell\|}{\sqrt{n}} - \sqrt{\bar{C}^{0,0}_{\phi,\ell}}\right|. \tag{55}$$

Since $p \geq 2$, Jensen's inequality bounds the p-th moment of the first term by $K_0^p e_\ell^p$. For the second, use the lower bound $\bar{C}^{0,0}_{\phi,\ell} \geq c > 0$ from Lemma D.1 and $|\sqrt{u} - \sqrt{v}| \leq |u - v|/\sqrt{v}$ to obtain

$$\left\|\frac{\|\bar{\phi}^0_\ell\|}{\sqrt{n}} - \sqrt{\bar{C}^{0,0}_{\phi,\ell}}\right\|_{L^p} \leq C\left\|\frac{1}{n}\sum_i \big((\bar{\phi}^0_{\ell,i})^2 - \bar{C}^{0,0}_{\phi,\ell}\big)\right\|_{L^p} \leq C_p n^{-1/2}.$$

The last inequality follows from Rosenthal's inequality: the summands are iid and centered, with uniformly bounded p-th moments by the L^{2p} feature bounds in Lemma D.1. Consequently,

$$\mathbb{E}\left|\sqrt{C^{0,0}_{\phi,\ell}} - \sqrt{\bar{C}^{0,0}_{\phi,\ell}}\right|^p \leq C_p e_\ell^p + C_p n^{-p/2}. \tag{56}$$

Subtracting the state recursions gives

$$h^0_{\ell,i} - \bar{h}^0_{\ell,i} = h^0_{\ell-1,i} - \bar{h}^0_{\ell-1,i} + \frac{1}{\sqrt{L}}\left(\sqrt{C^{0,0}_{\phi,\ell-1}} - \sqrt{\bar{C}^{0,0}_{\phi,\ell-1}}\right) z^0_{\ell,i}.$$

Conditional on the preceding forward history, the current state error and amplitude difference are fixed, while $z^0_{\ell,i}$ is standard Gaussian. Applying Eq. (23) with $\varepsilon = L^{-1/2}$, taking expectations, averaging over coordinates, and using Eq. (56) gives

$$e^p_\ell \le \left(1 + \frac{C_p}{L}\right) e^p_{\ell-1} + \frac{C_p}{L} n^{-p/2}, \qquad e_0 = 0.$$

Discrete Gronwall therefore gives

$$\sup_{L\ge 1} \max_{0\le \ell\le L} e_\ell \le C_p n^{-1/2}.$$

The coupling is exchangeable, so e_ℓ equals each coordinate error norm, proving the convergence estimate. Finally, Minkowski's inequality, Lemma D.1, and linear growth of ϕ give the uniform finite-width state and feature moments. □

Corollary D.3 (Empirical feature second moments)**.** *Suppose that $q_0 > 0$ and that ϕ is globally Lipschitz and not identically zero. For every $p \ge 2$,*

$$\sup_{L\ge 1} \sup_{0\le \ell\le L} \|C^{0,0,n,L}_{\phi,\ell} - \bar{C}^{0,0,L}_{\phi,\ell}\|_{L^p} \le C_p n^{-1/2}, \tag{57}$$

where C_p is independent of n, L.

Proof. Apply Lemma A.5 with $X_i = Y_i = \phi^{0,n,L}_{\ell,i}$ and $\bar{X}_i = \bar{Y}_i = \bar{\phi}^{0,L}_{\ell,i}$. The required L^{2p} bounds and coupling estimates follow from Lemma D.1 and the forward width estimate. □

D.3 Decoupling the Backward Process

Set

$$P_{\ell-1} := \frac{\phi^0_{\ell-1}(\phi^0_{\ell-1})^\top}{\|\phi^0_{\ell-1}\|^2}, \qquad D_{\ell-1} := \operatorname{diag}(\phi'(h^0_{\ell-1})). \tag{58}$$

When $\phi^0_{\ell-1} = 0$, set $P_{\ell-1} = 0$ and interpret all ratios involving $\|\phi^0_{\ell-1}\|$ as zero. The Gaussian representation from Proposition 4.1 admits a joint matrix realization

$$W^0_\ell = \sqrt{n}\, a^0_\ell \frac{(\phi^0_{\ell-1})^\top}{\|\phi^0_{\ell-1}\|^2} + \widetilde{W}_\ell (I - P_{\ell-1}), \tag{59}$$

where the matrices $\widetilde{W}_\ell$ have iid standard Gaussian entries, are independent of the forward trajectories and the readout weights. Define the auxiliary backward process by

$$\widetilde{g}^0_L = v^0, \qquad \widetilde{g}^0_{\ell-1} = \widetilde{g}^0_\ell + \frac{1}{\sqrt{L}} D_{\ell-1} \widetilde{b}^0_\ell, \qquad \widetilde{b}^0_\ell := \frac{\widetilde{W}_\ell^\top \widetilde{g}^0_\ell}{\sqrt{n}}. \tag{60}$$

Proposition D.4 (Backward decoupling)**.** *Suppose that $q_0 > 0$ and that ϕ is differentiable, globally Lipschitz, and not identically zero. For every $p \ge 2$, the joint realization above satisfies*

$$\sup_{L\ge 1} \sup_{0\le \ell\le L} \sup_{1\le i\le n} \|g^{0,n,L}_{\ell,i} - \widetilde{g}^{0,n,L}_{\ell,i}\|_{L^p} \le C_p n^{-1/2}.$$

Both backward processes have uniformly bounded coordinate moments of every finite p-th order.

Proof. We suppress n, L in the process notation. Let K_0 be a global Lipschitz constant for ϕ, so that $|\phi'| \le K_0$. The terminal case $\ell = L$ is immediate. Fix $0 \le \ell < L$, and a coordinate i. Starting from $y_\ell = \widetilde{y}_\ell = e_i$, define the tangent vectors forward

$$y_j = y_{j-1} + \frac{W^0_j D_{j-1} y_{j-1}}{\sqrt{Ln}}, \qquad \widetilde{y}_j = \widetilde{y}_{j-1} + \frac{\widetilde{W}_j D_{j-1} \widetilde{y}_{j-1}}{\sqrt{Ln}}, \qquad \ell < j \le L. \tag{61}$$

Since the terminal Gaussian vector v^0 is independent of both tangent paths,

$$g^0_{\ell,i} - \widetilde{g}^0_{\ell,i} = \langle y_L - \widetilde{y}_L, v^0 \rangle, \qquad \mathbb{E}|g^0_{\ell,i} - \widetilde{g}^0_{\ell,i}|^p \le C_p \mathbb{E}\|y_L - \widetilde{y}_L\|^p. \tag{62}$$

It therefore suffices to bound the tangent difference.

Let $e_j := y_j - \widetilde{y}_j$. The matrix realization Eq. (59) gives

$$e_j = e_{j-1} + \frac{1}{\sqrt{Ln}}\left[\widetilde{W}_j\big(D_{j-1}e_{j-1} - P_{j-1}D_{j-1}y_{j-1}\big) + \sqrt{n}\,a_j^0 \frac{(\phi_{j-1}^0)^\top D_{j-1}y_{j-1}}{\|\phi_{j-1}^0\|^2}\right]. \tag{63}$$

Condition only on the input h_0^0 and the variables $a_k^0, \widetilde{W}_k$ for $1 \le k < j$. The tangent vectors and feature direction are then all fixed. Under this conditioning, a_j^0 and $\widetilde{W}_j$ are independent and centered Gaussian variables. Thus, the bracket in Eq. (63) is centered Gaussian with covariance

$$\|D_{j-1}e_{j-1} - P_{j-1}D_{j-1}y_{j-1}\|^2 + \|P_{j-1}D_{j-1}y_{j-1}\|^2 I_n.$$

Since $y_{j-1} = \widetilde{y}_{j-1} + e_{j-1}$ and $\|D_{j-1}\|_{\mathrm{op}} \le K_0$, the factor is bounded by

$$C\|e_{j-1}\|^2 + C\|P_{j-1}D_{j-1}\widetilde{y}_{j-1}\|^2.$$

Lemma A.3, applied with $\varepsilon = L^{-1/2}$, therefore gives, after taking expectations,

$$\mathbb{E}\|e_j\|^p \le \left(1 + \frac{C_p}{L}\right)\mathbb{E}\|e_{j-1}\|^p + \frac{C_p}{L}\mathbb{E}\|P_{j-1}D_{j-1}\widetilde{y}_{j-1}\|^p. \tag{64}$$

It remains to bound the projection in the last term. For fixed $\ell < j \le L$, define

$$p_j := \frac{D_{j-1}\phi_{j-1}^0}{\|\phi_{j-1}^0\|}, \tag{65}$$

with $p_j = 0$ when $\phi_{j-1}^0 = 0$. Then $\|p_j\| \le K_0$ and

$$\|P_{j-1}D_{j-1}\widetilde{y}_{j-1}\| = |p_j^\top \widetilde{y}_{j-1}|.$$

Conditionally on the entire forward trajectory, p_j and all D_{k-1}, $1 \le k \le L$, are fixed, while the auxiliary matrices remain independent standard Gaussian matrices. Write $\mathbb{E}_{\widetilde{W}}$ for expectation over these matrices with the forward trajectory held fixed.

For each $\ell < k \le L$, given the preceding auxiliary matrices, the auxiliary tangent increment in Eq. (61) is centered Gaussian with covariance $\|D_{k-1}\widetilde{y}_{k-1}\|^2 I_n/(Ln)$. Therefore, Lemma A.3 and discrete Grönwall give

$$\sup_{\ell \le k \le L} \mathbb{E}_{\widetilde{W}}\|\widetilde{y}_k\|^p \le C_p. \tag{66}$$

Moreover, iterating the auxiliary tangent recursion gives

$$p_j^\top \widetilde{y}_{j-1} = p_{j,i} + \sum_{k=\ell+1}^{j-1} \frac{p_j^\top \widetilde{W}_k D_{k-1}\widetilde{y}_{k-1}}{\sqrt{Ln}}.$$

The partial sums form a martingale with conditionally centered Gaussian increments, whose variances are at most $K_0^4\|\widetilde{y}_{k-1}\|^2/(Ln)$. Lemma A.6 and Eq. (66) consequently imply

$$\mathbb{E}_{\widetilde{W}}|p_j^\top \widetilde{y}_{j-1}|^p \le C_p\big(|p_{j,i}|^p + n^{-p/2}\big).$$

To bound $p_{j,i}$, let $c > 0$ be the lower bound on $\bar{C}_{\phi,j-1}^{0,0}$ from Lemma D.1. On the event $C_{\phi,j-1}^{0,0} \ge c/2$, we have $\|\phi_{j-1}^0\|^2 = nC_{\phi,j-1}^{0,0} \ge nc/2$, and hence $|p_{j,i}|^p \le C_p n^{-p/2}|\phi_{j-1,i}^0|^p$. On the complementary event $C_{\phi,j-1}^{0,0} < c/2$, we have $|p_{j,i}| \le K_0$, while Markov's inequality and Eq. (57) give $\mathbb{P}(C_{\phi,j-1}^{0,0} < c/2) \le C_p n^{-p/2}$. Therefore, splitting the expectation over these two events gives

$$\mathbb{E}|p_{j,i}|^p \le C_p n^{-p/2}\mathbb{E}|\phi_{j-1,i}^0|^p + K_0^p\mathbb{P}(C_{\phi,j-1}^{0,0} < c/2) \le C_p n^{-p/2},$$

where the finite-width feature moments follow from the forward width estimate and Lemma D.1. Averaging the preceding martingale estimate over the forward trajectory yields

$$\sup_{\ell < j \le L} \mathbb{E}\|P_{j-1}D_{j-1}\widetilde{y}_{j-1}\|^p \le C_p n^{-p/2}. \tag{67}$$

Substituting this estimate into Eq. (64) gives

$$\mathbb{E}\|e_j\|^p \le \left(1 + \frac{C_p}{L}\right)\mathbb{E}\|e_{j-1}\|^p + \frac{C_p}{L}n^{-p/2}.$$

Since $e_\ell = 0$, discrete Gronwall and Eq. (62) prove the decoupling bound. The case $\ell = L$ is immediate. Finally, $\widetilde{g}_{\ell,i}^0 = \langle \widetilde{y}_L, v^0 \rangle$ and Eq. (66) give uniform auxiliary moments. The true backward moments follow by Minkowski's inequality. □

D.4 Width Convergence of the Backward Process

Proposition D.5 (Backward width convergence)**.** *Suppose that $q_0 > 0$ and that $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. The finite-width and mean-field processes admit a joint coupling such that, for every $p \geq 2$,*

$$\sup_{L\geq 1}\sup_{0\leq \ell\leq L}\sup_{1\leq i\leq n}\left(\|\widetilde{g}_{\ell,i}^{0,n,L}-\bar{g}_{\ell,i}^{0,L}\|_{L^p}+\|g_{\ell,i}^{0,n,L}-\bar{g}_{\ell,i}^{0,L}\|_{L^p}\right)\leq C_p n^{-1/2}. \tag{68}$$

In the same coupling,

$$\sup_{L\geq 1}\sup_{1\leq \ell\leq L}\sup_{1\leq i\leq n}\|\widetilde{b}_{\ell,i}^{0,n,L}-\zeta_{\ell,i}^{0,L}\|_{L^p}\leq C_p n^{-1/2}. \tag{69}$$

Proof. We suppress n, L in the process notation. Construct the coupling in reverse layer order, with $\bar{g}_L^0 = \widetilde{g}_L^0 = v^0$. Before layer ℓ, condition on the forward trajectories and the auxiliary matrices at layers greater than ℓ. Then $\widetilde{g}_\ell^0$ is fixed, while $\widetilde{W}_\ell$ remains independent standard Gaussian. Set

$$\widetilde{z}_\ell^0 := \frac{\widetilde{W}_\ell^\top \widetilde{g}_\ell^0}{\|\widetilde{g}_\ell^0\|}, \qquad \zeta_\ell^0 := \sqrt{\bar{C}_{g,\ell}^{0,0}}\,\widetilde{z}_\ell^0,$$

using an independent standard Gaussian vector when $\widetilde{g}_\ell^0 = 0$. These innovative Gaussian vectors are independent of the coupled forward trajectories. In this coupling, the innovative increment is

$$\widetilde{b}_\ell^0 = \frac{\|\widetilde{g}_\ell^0\|}{\sqrt{n}}\widetilde{z}_\ell^0.$$

The two backward recursions now use the same Gaussian innovations, so we consider

$$e_\ell := \left(\mathbb{E}\frac{1}{n}\sum_i |\widetilde{g}_{\ell,i}^0 - \bar{g}_{\ell,i}^0|^p\right)^{1/p}.$$

As in the forward width proof, the reverse triangle inequality gives

$$\left|\frac{\|\widetilde{g}_\ell^0\|}{\sqrt{n}} - \sqrt{\bar{C}_{g,\ell}^{0,0}}\right| \leq \frac{\|\widetilde{g}_\ell^0 - \bar{g}_\ell^0\|}{\sqrt{n}} + \left|\frac{\|\bar{g}_\ell^0\|}{\sqrt{n}} - \sqrt{\bar{C}_{g,\ell}^{0,0}}\right|.$$

Jensen's inequality bounds the first term in L^p by e_ℓ. Since the mean-field coordinates are iid, the uniform L^{2p} bounds and the lower bound $\bar{C}_{g,\ell}^{0,0} \geq 1$ from Lemma D.1, together with the square-root inequality and Lemma A.4, bound the second term by $C_p n^{-1/2}$. Hence

$$\left\|\frac{\|\widetilde{g}_\ell^0\|}{\sqrt{n}} - \sqrt{\bar{C}_{g,\ell}^{0,0}}\right\|_{L^p} \leq e_\ell + C_p n^{-1/2}. \tag{70}$$

Subtracting the backward recursions gives

$$\widetilde{g}_{\ell-1,i}^0 - \bar{g}_{\ell-1,i}^0 = \widetilde{g}_{\ell,i}^0 - \bar{g}_{\ell,i}^0 + \frac{1}{\sqrt{L}}\left[\phi'(h_{\ell-1,i}^0)\frac{\|\widetilde{g}_\ell^0\|}{\sqrt{n}} - \phi'(\bar{h}_{\ell-1,i}^0)\sqrt{\bar{C}_{g,\ell}^{0,0}}\right]\widetilde{z}_{\ell,i}^0.$$

Since ϕ' is bounded and globally Lipschitz, Eq. (70), the forward width estimate, and the uniform bound on $\bar{C}_{g,\ell}^{0,0}$ imply that the coefficient in brackets has L^p norm at most $C_p(e_\ell + n^{-1/2})$. Under the conditioning above, this coefficient and the current state error are fixed, while $\widetilde{z}_{\ell,i}^0$ is standard Gaussian. Lemma A.3, applied with $\varepsilon = L^{-1/2}$, therefore gives, after averaging,

$$e_{\ell-1}^p \leq \left(1 + \frac{C_p}{L}\right)e_\ell^p + \frac{C_p}{L}n^{-p/2}, \qquad e_L = 0.$$

Backward discrete Gronwall and exchangeability prove the auxiliary state estimate. Proposition D.4 and the triangle inequality give the true state estimate.

Finally,

$$\widetilde{b}_{\ell,i}^0 - \zeta_{\ell,i}^0 = \left(\frac{\|\widetilde{g}_\ell^0\|}{\sqrt{n}} - \sqrt{\bar{C}_{g,\ell}^{0,0}}\right)\widetilde{z}_{\ell,i}^0.$$

Conditional Gaussian moments, Eq. (70), and the proved state estimate give Eq. (69). □

Corollary D.6 (Matrix evaluations and empirical second moments)**.** *Under the conditions of Proposition D.5, for every $p \geq 2$,*

$$\sup_{L\geq 1}\sup_{1\leq \ell\leq L}\sup_{1\leq i\leq n}\left(\|a_{\ell,i}^{0,n,L}-\xi_{\ell,i}^{0,L}\|_{L^p}+\|b_{\ell,i}^{0,n,L}-\zeta_{\ell,i}^{0,L}\|_{L^p}\right)\leq C_p n^{-1/2}, \tag{71}$$

$$\sup_{L\geq 1}\sup_{0\leq \ell\leq L}\left(\|C_{\phi,\ell}^{0,0,n,L}-\bar{C}_{\phi,\ell}^{0,0,L}\|_{L^p}+\|C_{g,\ell}^{0,0,n,L}-\bar{C}_{g,\ell}^{0,0,L}\|_{L^p}\right)\leq C_p n^{-1/2}. \tag{72}$$

Proof. The covariance estimates follow from Eq. (25) and the state estimates. The forward evaluation estimate follows from $a_{\ell,i}^0-\xi_{\ell,i}^0=(\sqrt{C_{\phi,\ell-1}^{0,0}}-\sqrt{\bar{C}_{\phi,\ell-1}^{0,0}})z_{\ell,i}^0$ and Eq. (56).

For the backward evaluation, Eq. (59) gives

$$b_{\ell,i}^0-\zeta_{\ell,i}^0=(\sqrt{C_{g,\ell}^{0,0}}-\sqrt{\bar{C}_{g,\ell}^{0,0}})\widetilde{z}_{\ell,i}^0+\frac{\langle a_\ell^0,g_\ell^0\rangle}{\|\phi_{\ell-1}^0\|^2}\phi_{\ell-1,i}^0-\sqrt{C_{g,\ell}^{0,0}}(P_{\ell-1}\widetilde{z}_\ell^0)_i. \tag{73}$$

The first term has L^p norm $C_p n^{-1/2}$ by the covariance estimate, the positive mean-field variance, and independence of the current standard Gaussian vector from the conditioning history.

To bound the aligned term, start a true tangent path at $y_\ell = a_\ell^0$ and evolve $y_j = J_j y_{j-1}$ for $j > \ell$. Successive Gaussian increment estimates give $\mathbb{E}\|y_L\|^p \leq C_p\mathbb{E}\|a_\ell^0\|^p \leq C_p n^{p/2}$. Since $\langle a_\ell^0, g_\ell^0\rangle = \langle y_L, v^0\rangle$ and v^0 is independent of the path,

$$\|\langle a_\ell^0, g_\ell^0\rangle\|_{L^p}\leq C_p\sqrt{n}. \tag{74}$$

On $\Omega_\ell := \{C_{\phi,\ell-1}^{0,0}\geq c/2\}$, Hölder's inequality at order $2p$, the feature moments, and Eq. (74) bound the aligned term by $C_p n^{-1/2}$ in L^p. The projection term satisfies

$$\mathbb{E}\left[\mathbf{1}_{\Omega_\ell}|\sqrt{C_{g,\ell}^{0,0}}(P_{\ell-1}\widetilde{z}_\ell^0)_i|^p\right]\leq C_p n^{-p/2}\mathbb{E}[(C_{g,\ell}^{0,0})^{p/2}|\phi_{\ell-1,i}^0|^p]\leq C_p n^{-p/2}.$$

Here we conditioned on the forward trajectories and the backward history above layer ℓ, and then used uniform moments.

On Ω_ℓ^c, use the full matrix evaluation rather than the ratios in Eq. (73). Gaussian operator-norm moments and Hölder's inequality imply

$$\|b_{\ell,i}^0-\zeta_{\ell,i}^0\|_{L^{2p}}\leq \big\|\|W_\ell^0/\sqrt{n}\|_{\mathrm{op}}\big\|_{L^{4p}}\big\|\|g_\ell^0\|\big\|_{L^{4p}}+C_p\leq C_p\sqrt{n}.$$

Equation Eq. (57) gives $\mathbb{P}(\Omega_\ell^c)\leq C_p n^{-2p}$, by using its L^{4p} estimate. Another application of Hölder's inequality therefore gives

$$\|\mathbf{1}_{\Omega_\ell^c}(b_{\ell,i}^0-\zeta_{\ell,i}^0)\|_{L^p}\leq \mathbb{P}(\Omega_\ell^c)^{1/(2p)}\|b_{\ell,i}^0-\zeta_{\ell,i}^0\|_{L^{2p}}\leq C_p n^{-1/2}.$$

Combining the two events proves the backward evaluation estimate. □

Corollary D.7 (Initial output and loss derivative)**.** *Suppose that $q_0 > 0$ and that ϕ is globally Lipschitz and not identically zero. For the initialized output $f^{0,n,L} := n^{-1}\langle v^0, h_L^{0,n,L}\rangle$, the mean-field output is $\bar{f}^{0,L} = 0$. For every $p \geq 2$,*

$$\sup_{L\geq 1}\|f^{0,n,L}\|_{L^p}\leq C_p n^{-1/2}.$$

For a fixed label y, if $u\mapsto \partial_f\mathcal{L}(u,y)$ is globally Lipschitz, then, with $\chi^{0,n,L} := \partial_f\mathcal{L}(f^{0,n,L},y)$ and $\bar{\chi}^0 := \partial_f\mathcal{L}(0,y)$,

$$\sup_{L\geq 1}\|\chi^{0,n,L}-\bar{\chi}^0\|_{L^p}\leq C_p n^{-1/2}.$$

Proof. The initialized readout v^0 is independent of the forward trajectory. Conditional on that trajectory, $f^{0,n,L}$ is centered Gaussian with variance $\|h_L^{0,n,L}\|^2/n^2$. The forward coordinate moment bounds and Jensen's inequality give

$$\mathbb{E}|f^{0,n,L}|^p\leq\frac{C_p}{n^p}\mathbb{E}\|h_L^{0,n,L}\|^p\leq C_p n^{-p/2}.$$

The mean-field output is zero by centering and independence of the readout. The loss-derivative estimate follows from its Lipschitz continuity. □

D.5 Scalar Variance Equations and the Depth Limit

We now retain the discrete mean-field notation and suppress the coordinate index i. Define

$$q_\ell^L := \mathbb{E}[(\bar{h}_\ell^{0,L})^2], \qquad r_\ell^L := \mathbb{E}[(\bar{g}_\ell^{0,L})^2], \tag{75}$$

$$F(q) := \mathbb{E}[\phi(\sqrt{q}Z)^2], \qquad \mathcal{D}(q) := \mathbb{E}[\phi'(\sqrt{q}Z)^2], \qquad Z \sim \mathcal{N}(0,1). \tag{76}$$

Then $\bar{C}_{\phi,\ell}^{0,0,L} = F(q_\ell^L)$ and $\bar{C}_{g,\ell}^{0,0,L} = r_\ell^L$. The variance recursions are

$$q_0^L = q_0, \qquad q_\ell^L = q_{\ell-1}^L + \frac{1}{L}F(q_{\ell-1}^L), \tag{77}$$

$$r_L^L = 1, \qquad r_{\ell-1}^L = \left(1 + \frac{1}{L}\mathcal{D}(q_{\ell-1}^L)\right) r_\ell^L. \tag{78}$$

Write $t_\ell := \ell/L$.

Lemma D.8 (Scalar variance limits)**.** *Suppose that $q_0 > 0$ and that $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. The equations*

$$q'(t) = F(q(t)), \quad q(0) = q_0, \qquad r'(t) = -\mathcal{D}(q(t))r(t), \quad r(1) = 1 \tag{79}$$

have unique solutions on $[0,1]$. There is $Q < \infty$ such that $q(t), q_\ell^L \in [q_0, Q]$. On this interval, $F, \sqrt{F}$, and $\mathcal{D}$ are Lipschitz, and F is bounded below by a positive constant. Moreover,

$$r(t) = \exp\left(\int_t^1 \mathcal{D}(q(s))\, ds\right), \qquad 1 \le r(t), r_\ell^L \le e^{K_0^2},$$

and

$$\max_{0\le\ell\le L} \left(|q_\ell^L - q(t_\ell)| + |r_\ell^L - r(t_\ell)|\right) \le \frac{C}{L}. \tag{80}$$

Proof. Lipschitz continuity of ϕ gives linear growth $F(q) \le 2K_0^2 q + 2|\phi(0)|^2$. Continuous and discrete Gronwall inequalities therefore give $q_0 \le q(t), q_\ell^L \le Q$, uniformly in $t \in [0,1]$, $L \ge 1$, and $0 \le \ell \le L$. On this common interval $[q_0, Q]$,

$$|F(q) - F(\widehat{q})| \le K_0|\sqrt{q} - \sqrt{\widehat{q}}|\, \mathbb{E}\left[|Z|\big(|\phi(\sqrt{q}Z)| + |\phi(\sqrt{\widehat{q}}Z)|\big)\right] \le C|q - \widehat{q}|,$$

$$|\mathcal{D}(q) - \mathcal{D}(\widehat{q})| \le 2K_0K_1\mathbb{E}|Z|\,|\sqrt{q} - \sqrt{\widehat{q}}| \le C|q - \widehat{q}|.$$

Since F is locally Lipschitz on $(0,\infty)$, the Picard–Lindelöf theorem gives a unique local solution q. The linear-growth bound and $q(t) \ge q_0$ allow this solution to extend uniquely to $[0,1]$. On the common variance interval $[q_0, Q]$, F is Lipschitz and has a strictly positive minimum, so $\sqrt{F}$ is also Lipschitz there. The explicit formula for r gives its existence, uniqueness, and bounds. The discrete backward bounds follow from Eq. (78).

By the Euler local truncation estimates in Lemma A.8, applied forward to $q' = F(q)$ and backward to $r' = -\mathcal{D}(q)r$, with $\Delta := 1/L$,

$$q(t_\ell) = q(t_{\ell-1}) + \Delta F(q(t_{\ell-1})) + O(\Delta^2),$$

$$r(t_{\ell-1}) = \big(1 + \Delta\mathcal{D}(q(t_{\ell-1}))\big)r(t_\ell) + O(\Delta^2),$$

uniformly in ℓ, L. Subtracting Eq. (77) from the first expansion and applying discrete Gronwall gives $\max_\ell |q_\ell^L - q(t_\ell)| \le C\Delta$. For the backward variance, set $e_\ell := |r_\ell^L - r(t_\ell)|$. Since $\mathcal{D} \le K_0^2$ and $\mathcal{D}$ is Lipschitz on $[q_0, Q]$, the second expansion and Eq. (78) give

$$e_{\ell-1} \le (1 + K_0^2\Delta)e_\ell + C\Delta|q_{\ell-1}^L - q(t_{\ell-1})| + C\Delta^2, \qquad e_L = 0.$$

The forward variance estimate bounds the last two terms by $C\Delta^2$. Backward discrete Gronwall therefore gives $\max_\ell e_\ell \le C\Delta$. □

Proposition D.9 (Well-posedness of the limiting system)**.** *Suppose that $q_0 > 0$ and that $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. Let $H_0^0 \sim \mathcal{N}(0, q_0)$, $v^0 \sim \mathcal{N}(0,1)$, and standard Brownian motions $B^0, \widetilde{B}$ be mutually independent. The system*

$$H_t^0 = H_0^0 + \int_0^t \sqrt{\mathbb{E}[\phi(H_s^0)^2]}\, dB_s^0, \tag{81}$$

$$G_t^0 = v^0 + \int_t^1 \phi'(H_s^0)\sqrt{\mathbb{E}[(G_s^0)^2]}\, d\widetilde{B}_s \tag{82}$$

has a unique continuous solution with finite second moments of the path supremum, relative to these prescribed primitives. The backward integral is an ordinary Itô integral after revealing the independent forward path at time zero. Explicitly,

$$H_t^0 = H_0^0 + \int_0^t \sqrt{F(q(s))}\, dB_s^0, \qquad G_t^0 = v^0 + \int_t^1 \phi'(H_s^0)\sqrt{r(s)}\, d\widetilde{B}_s, \tag{83}$$

where q, r solve Eq. (79). Their second moments are $q(t)$ and $r(t)$, respectively. For every $p \geq 2$,

$$\left\| \sup_{0\leq t\leq 1} |H_t^0| \right\|_{L^p} + \left\| \sup_{0\leq t\leq 1} |G_t^0| \right\|_{L^p} \leq C_p, \tag{84}$$

$$\|H_t^0 - H_s^0\|_{L^p} + \|G_t^0 - G_s^0\|_{L^p} \leq C_p |t-s|^{1/2}. \tag{85}$$

Proof. Construct the processes by Eq. (83). The forward process is centered Gaussian with variance $q_0 + \int_0^t F(q(s))\, ds = q(t)$, so its coefficient has the required self-consistent form. Since the full forward path and v^0 are independent of $\widetilde{B}$, revealing them at time zero preserves the Brownian property of $\widetilde{B}$. The backward integrand is then bounded and predictable. Itô isometry gives

$$\mathbb{E}[(G_t^0)^2] = 1 + \int_t^1 \mathcal{D}(q(s))r(s)\, ds = r(t).$$

This verifies both equations.

For any solution, the forward diffusion coefficient is deterministic, so the forward field is centered Gaussian. Its variance therefore solves the first equation in Eq. (79). Uniqueness of q fixes the forward coefficient and, with the same primitives, the forward path. The backward Itô isometry then shows that the backward variance solves the second equation. Uniqueness of r fixes the backward coefficient and path. This proves pathwise uniqueness. The explicit representations also give uniqueness in law.

Since $q(s) \in [q_0, Q]$, $r(s) \leq e^{K_0^2}$, and $|\phi'| \leq K_0$, both diffusion coefficients are uniformly bounded. For the forward process, Lemma A.9 gives

$$\left\| \sup_{0\leq t\leq 1} |H_t^0| \right\|_{L^p} \leq \|H_0^0\|_{L^p} + C_p \left(\int_0^1 F(q(s))\, ds \right)^{1/2} \leq C_p,$$

since H_0^0 is Gaussian and F is bounded on $[q_0, Q]$.

For the backward process, reveal the entire independent forward path and define the Itô martingale

$$M_t := \int_0^t \phi'(H_s^0)\sqrt{r(s)}\, d\widetilde{B}_s.$$

Lemma A.9 and the coefficient bounds give

$$\left\| \sup_{0\leq t\leq 1} |M_t| \right\|_{L^p} \leq C_p \left\| \left(\int_0^1 \phi'(H_s^0)^2 r(s)\, ds \right)^{1/2} \right\|_{L^p} \leq C_p.$$

Since $G_t^0 = v^0 + M_1 - M_t$ and v^0 is Gaussian,

$$\left\| \sup_{0\leq t\leq 1} |G_t^0| \right\|_{L^p} \leq \|v^0\|_{L^p} + 2 \left\| \sup_{0\leq t\leq 1} |M_t| \right\|_{L^p} \leq C_p.$$

Together, these estimates prove Eq. (84).

For the temporal increments, let $0 \le s \le t \le 1$. The explicit representations give

$$H_t^0 - H_s^0 = \int_s^t \sqrt{F(q(u))}\, dB_u^0, \qquad G_t^0 - G_s^0 = -\int_s^t \phi'(H_u^0)\sqrt{r(u)}\, d\widetilde{B}_u.$$

Applying Lemma A.9 over $[s,t]$ yields

$$\|H_t^0 - H_s^0\|_{L^p} \le C_p \left(\int_s^t F(q(u))\, du\right)^{1/2} \le C_p|t-s|^{1/2},$$
$$\|G_t^0 - G_s^0\|_{L^p} \le C_p \left\| \left(\int_s^t \phi'(H_u^0)^2 r(u)\, du\right)^{1/2} \right\|_{L^p} \le C_p|t-s|^{1/2}.$$

This proves Eq. (85).

□

D.6 Depth Convergence and the Combined Approximation

Couple the discrete mean fields and the continuous limit using the same $H_0^0, v^0, B^0, \widetilde{B}$, with

$$\bar{h}_\ell^{0,L} = \bar{h}_{\ell-1}^{0,L} + \sqrt{F(q_{\ell-1}^L)}(B_{t_\ell}^0 - B_{t_{\ell-1}}^0), \tag{86}$$
$$\bar{g}_{\ell-1}^{0,L} = \bar{g}_\ell^{0,L} + \phi'(\bar{h}_{\ell-1}^{0,L})\sqrt{r_\ell^L}(\widetilde{B}_{t_\ell} - \widetilde{B}_{t_{\ell-1}}). \tag{87}$$

For $t \in [t_{\ell-1}, t_\ell]$, define the continuous stochastic interpolations by

$$\bar{h}^{0,L}(t) := \bar{h}_{\ell-1}^{0,L} + \sqrt{F(q_{\ell-1}^L)}(B_t^0 - B_{t_{\ell-1}}^0), \tag{88}$$
$$\bar{g}^{0,L}(t) := \bar{g}_\ell^{0,L} + \phi'(\bar{h}_{\ell-1}^{0,L})\sqrt{r_\ell^L}(\widetilde{B}_{t_\ell} - \widetilde{B}_t). \tag{89}$$

These definitions agree at the cell endpoints and recover the mean-field states at every t_ℓ.

Proposition D.10 (Strong depth convergence)**.** *Suppose that $q_0 > 0$ and that $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. For every $p \ge 2$, the preceding coupling satisfies*

$$\left\| \sup_{0\le t\le 1} |\bar{h}^{0,L}(t) - H_t^0| \right\|_{L^p} \le C_p L^{-1}, \tag{90}$$
$$\left\| \sup_{0\le t\le 1} |\bar{g}^{0,L}(t) - G_t^0| \right\|_{L^p} \le C_p L^{-1/2}. \tag{91}$$

The joint laws of the stochastic interpolations converge to the law of (H^0, G^0) in p-Wasserstein distance on $C([0,1];\mathbb{R}^2)$ at rate $C_p L^{-1/2}$, with the forward marginal converging at rate $C_p L^{-1}$.

Proof. For $s \in (t_{\ell-1}, t_\ell]$, Eq. (80) and the Lipschitz continuity of q, r give

$$|q_{\ell-1}^L - q(s)| + |r_\ell^L - r(s)| \le C/L.$$

Since $\sqrt{F}$ is Lipschitz on $[q_0, Q]$, the forward integrand difference is bounded by C/L. The shared Brownian coupling and Lemma A.9 therefore give

$$\left\| \sup_t |\bar{h}^{0,L}(t) - H_t^0| \right\|_{L^p} \le C_p \left(\sum_{\ell=1}^L \int_{t_{\ell-1}}^{t_\ell} \left| \sqrt{F(q_{\ell-1}^L)} - \sqrt{F(q(s))} \right|^2 ds \right)^{1/2} \le C_p L^{-1},$$

proving Eq. (90).

For the backward coefficient, the forward state is evaluated at the left grid point. The forward depth bound and Eq. (85) give

$$\begin{aligned} \|\bar{h}_{\ell-1}^{0,L} - H_s^0\|_{L^p} &\le \|\bar{h}_{\ell-1}^{0,L} - H_{t_{\ell-1}}^0\|_{L^p} + \|H_{t_{\ell-1}}^0 - H_s^0\|_{L^p} \\ &\le C_p \left(L^{-1} + |s - t_{\ell-1}|^{1/2} \right) \le C_p L^{-1/2}. \end{aligned}$$

Since ϕ' is bounded and Lipschitz and $1 \le r(s), r_\ell^L \le e^{K_0^2}$,

$$\|\phi'(\bar{h}_{\ell-1}^{0,L})\sqrt{r_\ell^L} - \phi'(H_s^0)\sqrt{r(s)}\|_{L^p} \le C\|\bar{h}_{\ell-1}^{0,L} - H_s^0\|_{L^p} + C|r_\ell^L - r(s)| \le C_p L^{-1/2}.$$

After revealing the independent forward primitives, the backward error is a tail Itô integral, whose supremum is bounded by twice that of the corresponding martingale. Lemma A.9 and Minkowski's inequality thus give

$$\left\| \sup_t |\bar{g}^{0,L}(t) - G_t^0| \right\|_{L^p} \le C_p \left(\int_0^1 L^{-1}\, ds \right)^{1/2} = C_p L^{-1/2}.$$

This proves Eq. (91). The Wasserstein estimates follow from the same coupling. □

Proposition D.11 (Combined width and depth approximation)**.** *Suppose that $q_0 > 0$ and that $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. For every $p \ge 2$ and every n, L, there is a joint coupling of the finite-width states and independent coordinate copies $(H_{\cdot,i}^0, G_{\cdot,i}^0)$ of the limiting pair such that*

$$\sup_{0\le\ell\le L} \sup_{1\le i\le n} \|h_{\ell,i}^{0,n,L} - H_{t_\ell,i}^0\|_{L^p} \le C_p(n^{-1/2} + L^{-1}), \tag{92}$$

$$\sup_{0\le\ell\le L} \sup_{1\le i\le n} \|g_{\ell,i}^{0,n,L} - G_{t_\ell,i}^0\|_{L^p} \le C_p(n^{-1/2} + L^{-1/2}). \tag{93}$$

Proof. The width coupling and the Brownian depth coupling have the same joint law for the discrete mean fields. They therefore admit a joint extension over these mean fields. Apply Minkowski's inequality to

$$h_{\ell,i}^{0,n,L} - H_{t_\ell,i}^0 = (h_{\ell,i}^{0,n,L} - \bar{h}_{\ell,i}^{0,L}) + (\bar{h}_{\ell,i}^{0,L} - H_{t_\ell,i}^0)$$

and

$$g_{\ell,i}^{0,n,L} - G_{t_\ell,i}^0 = (g_{\ell,i}^{0,n,L} - \widetilde{g}_{\ell,i}^{0,n,L}) + (\widetilde{g}_{\ell,i}^{0,n,L} - \bar{g}_{\ell,i}^{0,L}) + (\bar{g}_{\ell,i}^{0,L} - G_{t_\ell,i}^0).$$

The established width and depth bounds give both assertions. □

The joint pairs $(H_{\cdot,i}^0, G_{\cdot,i}^0)$ are independent across width coordinates, although $H_{\cdot,i}^0$ and $G_{\cdot,i}^0$ generally remain dependent within each pair. The combined estimates provide the initialization-time forward–backward characterization in Theorem 4.3.

E Convergence Proofs After One SGD Update

In this section, we consider one gradient step with $\eta = \eta_c n$ for a fixed $\eta_c > 0$ following μP. We analyze the updated forward pass and readout using the initialization backward process. Write $q_0 = \|x^0\|^2/d > 0$ and, for $k, j \in \{0, 1\}$, define

$$q_x^{k,j} := \frac{\langle x^k, x^j \rangle}{d}. \tag{94}$$

Here x^1 is the evaluation input, not another update. Throughout, ϕ is assumed to lie in $C^1(\mathbb{R})$, not be identically zero, and have Lipschitz ϕ and ϕ'. We state these hypotheses again in the formal results.

E.1 Updated Network and Discrete Mean Field

Write $\chi^0 = \partial_f \mathcal{L}(f^0, y)$ and $\bar{\chi}^0 = \partial_f \mathcal{L}(0, y)$. The gradients at the training input x^0 are

$$\nabla_U \mathcal{L} = \frac{\chi^0}{n\sqrt{d}} g_0^0(x^0)^\top, \qquad \nabla_{W_\ell} \mathcal{L} = \frac{\chi^0}{n\sqrt{Ln}} g_\ell^0(\phi_{\ell-1}^0)^\top, \qquad \nabla_v \mathcal{L} = \frac{\chi^0}{n} h_L^0. \tag{95}$$

A gradient step with learning rate $\eta = \eta_c n$ therefore gives

$$U^1 = U^0 - \frac{\eta_c \chi^0}{\sqrt{d}} g_0^0(x^0)^\top, \quad W_\ell^1 = W_\ell^0 - \frac{\eta_c \chi^0}{\sqrt{Ln}} g_\ell^0(\phi_{\ell-1}^0)^\top, \quad v^1 = v^0 - \eta_c \chi^0 h_L^0. \tag{96}$$

Evaluating the updated network at x^1 yields the estimate $f^1 = \frac{1}{n}\langle v^1, h_L^1\rangle$ with

$$h_0^1 = \frac{1}{\sqrt{d}} U^0 x^1 - \eta_c \chi^0 q_x^{1,0} g_0^0, \tag{97}$$

$$h_\ell^1 = h_{\ell-1}^1 + \frac{1}{\sqrt{L}} a_\ell^1 - \frac{\eta_c \chi^0}{L} C_{\phi,\ell-1}^{1,0} g_\ell^0, \tag{98}$$

$$a_\ell^1 = \frac{1}{\sqrt{n}} W_\ell^0 \phi_{\ell-1}^1, \tag{99}$$

where $\phi_{\ell,i}^1 = \phi(h_{\ell,i}^1)$. For $k, j \in \{0, 1\}$, define the empirical and mean-field kernels

$$C_{\theta,\ell}^{k,j,n,L} := \frac{1}{n}\langle \theta_\ell^{k,n,L}, \theta_\ell^{j,n,L}\rangle, \qquad \bar{C}_{\theta,\ell}^{k,j,L} := \mathbb{E}[\bar{\theta}_{\ell,i}^{k,L} \bar{\theta}_{\ell,i}^{j,L}]. \tag{100}$$

The choices of θ used below are activated features and initial backward states.

The reuse formula from Corollary C.4 identifies the Gaussian and response contributions to the limit. At the input layer, define the mean field variable for the initial state

$$\bar{h}_{0,i}^1 = \xi_{0,i}^1 - \eta_c \bar{\chi}^0 q_x^{1,0} \bar{g}_{0,i}^0, \qquad \xi_{0,i}^1 = \left(\frac{1}{\sqrt{d}} U^0 x^1\right)_i. \tag{101}$$

The initial forward construction has $\bar{h}_{0,i}^0 = \xi_{0,i}^0 = (U^0 x^0/\sqrt{d})_i$, so

$$\mathbb{E}[(\xi_{0,i}^1)^2] = \frac{\|x^1\|^2}{d}, \qquad \mathbb{E}[\xi_{0,i}^1 \xi_{0,i}^0] = q_x^{1,0}. \tag{102}$$

Recursively, once $\bar{h}_{\ell-1,i}^1$ is defined, set $\bar{\phi}_{\ell-1,i}^1 = \phi(\bar{h}_{\ell-1,i}^1)$ and define

$$\bar{\phi}_{\ell-1,i,\perp}^1 := \bar{\phi}_{\ell-1,i}^1 - \frac{\bar{C}_{\phi,\ell-1}^{1,0}}{\bar{C}_{\phi,\ell-1}^{0,0}} \bar{\phi}_{\ell-1,i}^0, \quad \xi_{\ell,i}^1 := \frac{\bar{C}_{\phi,\ell-1}^{1,0}}{\bar{C}_{\phi,\ell-1}^{0,0}} \xi_{\ell,i}^0 + \sqrt{\mathbb{E}[(\bar{\phi}_{\ell-1,i,\perp}^1)^2]}\, z_{\ell,i}^1. \tag{103}$$

Thus $(\xi_{\ell,i}^0, \xi_{\ell,i}^1)$ is a centered Gaussian pair with

$$\mathbb{E}[\xi_{\ell,i}^s \xi_{\ell,i}^t] = \bar{C}_{\phi,\ell-1}^{s,t}, \quad s, t \in \{0, 1\}. \tag{104}$$

These pairs are independent across layers and coordinates, and independent of the initial backward Gaussian innovations $\zeta_{\ell,i}^0$ and v^0.

Define the the mean-field response by

$$\bar{R}_\ell^{1,0} := \frac{\sqrt{L}}{\bar{C}_{g,\ell}^{0,0}} \mathbb{E}[\zeta_{\ell,i}^0 \bar{\phi}_{\ell-1,i}^1] \tag{105}$$

The limiting updated recursion is therefore

$$\bar{h}_{\ell,i}^1 = \bar{h}_{\ell-1,i}^1 + \frac{1}{\sqrt{L}} \xi_{\ell,i}^1 + \frac{1}{L}\left(\bar{R}_\ell^{1,0} - \eta_c \bar{\chi}^0 \bar{C}_{\phi,\ell-1}^{1,0}\right) \bar{g}_{\ell,i}^0, \qquad 1 \le \ell \le L. \tag{106}$$

Lemma D.1 ensures that all denominators in these definitions are strictly positive.

E.2 Width Convergence of the Updated Forward Pass

The formal results use L^p norms with $p \ge 2$. In the width proofs, we verify the bounds at orders $2p$ with $p \ge 1$; these estimates imply the stated bounds at every L^q order $q \ge 2$.

Lemma E.1 (Moments of the limiting updated process)**.** *Fix $L \ge 1$, deterministic inputs $x^0, x^1 \in \mathbb{R}^d$ with $x^0 \ne 0$, and $\eta_c \ge 0$. Suppose that $q_0 > 0$ and $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. For a fixed label y, suppose that $\bar{\chi}^0 = \partial_f \mathcal{L}(0, y)$ exists and is finite. Then the limiting process defined by Eq. (101)–Eq. (106) satisfies, for every $p \ge 2$,*

$$\max_{0 \le \ell \le L} \sup_i \|\bar{h}_{\ell,i}^{1,L}\|_{L^p} \le C_{p,L}. \tag{107}$$

The joint limiting coordinate processes are iid across i.

Proof. The initial limiting states have all finite moments, so Eq. (101) gives the claim at layer 0. Suppose it holds at layer $\ell - 1$. The Lipschitz growth of ϕ and Cauchy–Schwarz give

$$|\bar{C}^{1,0}_{\phi,\ell-1}| \le \|\bar{\phi}^1_{\ell-1,i}\|_{L^2}\|\bar{\phi}^0_{\ell-1,i}\|_{L^2}, \qquad |\bar{R}^{1,0}_\ell| \le \frac{\sqrt{L}}{\sqrt{\bar{C}^{0,0}_{g,\ell}}}\|\bar{\phi}^1_{\ell-1,i}\|_{L^2}. \tag{108}$$

The second bound follows directly from the expectation definition of the response in Eq. (105). All coefficients and Gaussian variances at layer ℓ are finite, so Eq. (106) gives all finite moments at that layer. Induction proves Eq. (107).

Finally, each limiting coordinate process is the same function of an independent copy of the primitive Gaussian family. This proves the iid assertion. □

Remark E.2. The fixed-depth bound $C_{p,L}$ suffices for the width analysis. By closing the response recursion, Lemma E.11 below strengthens it to a bound C_p independent of L.

Lemma E.3 (Approximation of the reused weight action)**.** *Fix $L \ge 1$ and $1 \le \ell \le L$. Let $x^0, x^1 \in \mathbb{R}^d$ be deterministic with $x^0 \ne 0$. Suppose that $q_0 > 0$ and $\phi \in C^1(\mathbb{R})$ is not identically zero, with ϕ and ϕ' globally Lipschitz. Fix $\eta_c \ge 0$ and a label y, and suppose that $u \mapsto \mathcal{L}(u, y)$ is differentiable. Use the initial joint coupling constructed in Proposition D.5 and the fresh Gaussian variables of Corollary C.4. Suppose also that, for every $q \ge 1$,*

$$\sup_i \|h^{1,n,L}_{\ell-1,i} - \bar{h}^{1,L}_{\ell-1,i}\|_{L^q} \le C_{q,L}n^{-1/2}, \qquad \sup_i \|h^{1,n,L}_{\ell-1,i}\|_{L^q} \le C_{q,L}. \tag{109}$$

Then, for every $p \ge 2$,

$$\max_{s,t\in\{0,1\}} \|C^{s,t,n,L}_{\phi,\ell-1} - \bar{C}^{s,t,L}_{\phi,\ell-1}\|_{L^p} \le C_{p,L}n^{-1/2}, \tag{110}$$

$$\sup_i \left\| a^{1,n,L}_{\ell,i} - \xi^{1,L}_{\ell,i} - \frac{\bar{R}^{1,0,L}_\ell}{\sqrt{L}}\bar{g}^{0,L}_{\ell,i} \right\|_{L^p} \le C_{p,L}n^{-1/2}. \tag{111}$$

Proof. The Lipschitz property of ϕ transfers the induction hypothesis in Eq. (109) to the updated features. Together with the initial forward convergence in Proposition D.2, the empirical-product estimate Eq. (25) therefore proves Eq. (110).

Lemma D.1 gives uniform positive lower bounds on $\bar{C}^{0,0}_{\phi,\ell-1}$ and $\bar{C}^{0,0}_{g,\ell}$. To control the empirical denominators, define

$$\Omega_\ell := \left\{ C^{0,0}_{\phi,\ell-1} \ge \tfrac{1}{2}\bar{C}^{0,0}_{\phi,\ell-1}, \quad C^{0,0}_{g,\ell} \ge \tfrac{1}{2}\bar{C}^{0,0}_{g,\ell} \right\}. \tag{112}$$

These lower bounds, Markov's inequality, and Eq. (72) give

$$\mathbb{P}(\Omega_\ell^c) \le C_r n^{-r}, \qquad r \ge 1. \tag{113}$$

We estimate the error on Ω_ℓ and use this probability bound to control its complement.

To prove Eq. (111), we use Eq. (46) and Eq. (103) to split the error into three terms:

$$\begin{aligned} a^1_{\ell,i} - \xi^1_{\ell,i} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}}\bar{g}^0_{\ell,i} = &\left[\frac{C^{1,0}_{\phi,\ell-1}}{C^{0,0}_{\phi,\ell-1}}a^0_{\ell,i} - \frac{\bar{C}^{1,0}_{\phi,\ell-1}}{\bar{C}^{0,0}_{\phi,\ell-1}}\xi^0_{\ell,i}\right] \\ &+ \left[\frac{\langle b^0_\ell, \phi^1_{\ell-1,\perp}\rangle}{\|g^0_\ell\|^2}g^0_{\ell,i} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}}\bar{g}^0_{\ell,i}\right] \\ &+ \left[\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}}\big((I - P^0_{g,\ell})z^1_\ell\big)_i - \sqrt{\mathbb{E}[(\bar{\phi}^1_{\ell-1,i,\perp})^2]}\, z^1_{\ell,i}\right]. \end{aligned} \tag{114}$$

We bound each bracket on Ω_ℓ and control the complementary event.

For the first bracket, the denominator lower bounds on Ω_ℓ, Lemma E.1, and Eq. (110) give

$$\left\| \mathbf{1}_{\Omega_\ell}\left(\frac{C^{1,0}_{\phi,\ell-1}}{C^{0,0}_{\phi,\ell-1}} - \frac{\bar{C}^{1,0}_{\phi,\ell-1}}{\bar{C}^{0,0}_{\phi,\ell-1}}\right)\right\|_{L^{4p}} \le C_{p,L}n^{-1/2}.$$

Adding and subtracting $(\bar{C}^{1,0}_{\phi,\ell-1}/\bar{C}^{0,0}_{\phi,\ell-1})a^0_{\ell,i}$, Hölder's inequality and the initial evaluation bounds in Corollary D.6 therefore give

$$\sup_i \left\| \mathbf{1}_{\Omega_\ell} \left[\frac{C^{1,0}_{\phi,\ell-1}}{C^{0,0}_{\phi,\ell-1}} a^0_{\ell,i} - \frac{\bar{C}^{1,0}_{\phi,\ell-1}}{\bar{C}^{0,0}_{\phi,\ell-1}} \xi^0_{\ell,i} \right] \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{115}$$

We also need convergence of the residual features for the next two brackets. Their difference is

$$\phi^1_{\ell-1,i,\perp} - \bar{\phi}^1_{\ell-1,i,\perp} = \phi^1_{\ell-1,i} - \bar{\phi}^1_{\ell-1,i} - \left(\frac{C^{1,0}_{\phi,\ell-1}}{C^{0,0}_{\phi,\ell-1}} - \frac{\bar{C}^{1,0}_{\phi,\ell-1}}{\bar{C}^{0,0}_{\phi,\ell-1}} \right) \phi^0_{\ell-1,i} - \frac{\bar{C}^{1,0}_{\phi,\ell-1}}{\bar{C}^{0,0}_{\phi,\ell-1}} (\phi^0_{\ell-1,i} - \bar{\phi}^0_{\ell-1,i}).$$

The same covariance estimates, the induction hypothesis, and the initial feature convergence give

$$\sup_i \left\| \mathbf{1}_{\Omega_\ell} (\phi^1_{\ell-1,i,\perp} - \bar{\phi}^1_{\ell-1,i,\perp}) \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}, \qquad \sup_i \| \mathbf{1}_{\Omega_\ell} \phi^1_{\ell-1,i,\perp} \|_{L^{2p}} \le C_{p,L}. \tag{116}$$

Here Hölder's inequality uses the covariance and feature estimates at order $4p$. The second bound follows by adding the uniformly bounded limiting residual.

For the second bracket, we first estimate the empirical numerator using the product decomposition from Lemma A.5. On Ω_ℓ, Hölder's inequality at doubled moment order, Eq. (71), and Eq. (116) control the coupled product differences. The limiting pairs $(\zeta^0_{\ell,i}, \bar{\phi}^1_{\ell-1,i,\perp})$ are iid across i by Lemma E.1. Lemma A.4 therefore controls the remaining centered average unconditionally; multiplication by $\mathbf{1}_{\Omega_\ell}$ only decreases its norm. Thus,

$$\left\| \mathbf{1}_{\Omega_\ell} \left[\frac{1}{n} \langle b^0_\ell, \phi^1_{\ell-1,\perp} \rangle - \mathbb{E}[\zeta^0_{\ell,i} \bar{\phi}^1_{\ell-1,i,\perp}] \right] \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{117}$$

Since $\zeta^0_{\ell,i}$ is centered and independent of $\bar{\phi}^0_{\ell-1,i}$, Eq. (105) gives

$$\mathbb{E}[\zeta^0_{\ell,i} \bar{\phi}^1_{\ell-1,i,\perp}] = \mathbb{E}[\zeta^0_{\ell,i} \bar{\phi}^1_{\ell-1,i}] = \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}} \bar{C}^{0,0}_{g,\ell}.$$

On Ω_ℓ, the denominator satisfies $C^{0,0}_{g,\ell} \ge \bar{C}^{0,0}_{g,\ell}/2 \ge 1/2$. Moreover, Lemmas D.1 and E.1 bound $\bar{R}^{1,0}_\ell/\sqrt{L}$. Applying Eq. (117) and Eq. (72), both at order $4p$, therefore gives

$$\left\| \mathbf{1}_{\Omega_\ell} \left[\frac{\langle b^0_\ell, \phi^1_{\ell-1,\perp} \rangle}{\|g^0_\ell\|^2} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}} \right] \right\|_{L^{4p}} \le C_{p,L} n^{-1/2}.$$

Adding and subtracting $(\bar{R}^{1,0}_\ell/\sqrt{L}) g^0_{\ell,i}$, Hölder's inequality, the uniform backward moments, and Proposition D.5 yield

$$\sup_i \left\| \mathbf{1}_{\Omega_\ell} \left[\frac{\langle b^0_\ell, \phi^1_{\ell-1,\perp} \rangle}{\|g^0_\ell\|^2} g^0_{\ell,i} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}} \bar{g}^0_{\ell,i} \right] \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{118}$$

For the third bracket, write $\bar{s}_\ell := \sqrt{\mathbb{E}[(\bar{\phi}^1_{\ell-1,i,\perp})^2]}$. We separate the Gaussian coefficient error from the projection correction:

$$\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} \big((I - P^0_{g,\ell}) z^1_\ell \big)_i - \bar{s}_\ell z^1_{\ell,i} = \left(\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right) z^1_{\ell,i} - \frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} (P^0_{g,\ell} z^1_\ell)_i. \tag{119}$$

We insert $n^{-1/2} \|\bar{\phi}^1_{\ell-1,\perp}\|$ and apply the triangle and reverse triangle inequalities:

$$\left| \frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right| \le \frac{\|\phi^1_{\ell-1,\perp} - \bar{\phi}^1_{\ell-1,\perp}\|}{\sqrt{n}} + \left| \frac{\|\bar{\phi}^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right|. \tag{120}$$

For the first term, Jensen's inequality gives

$$\left\| \mathbf{1}_{\Omega_\ell} \frac{\|\phi^1_{\ell-1,\perp} - \bar{\phi}^1_{\ell-1,\perp}\|}{\sqrt{n}} \right\|^{2p}_{L^{2p}} \le \frac{1}{n} \sum_i \mathbb{E} \left[\mathbf{1}_{\Omega_\ell} |\phi^1_{\ell-1,i,\perp} - \bar{\phi}^1_{\ell-1,i,\perp}|^{2p} \right] \le C_{p,L} n^{-p}, \tag{121}$$

where the last inequality uses Eq. (116).

For the second term, if $\bar{s}_\ell > 0$, using $|\sqrt{u} - \sqrt{v}| \le |u - v|/\sqrt{v}$ gives

$$\left| \frac{\|\bar{\phi}^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right| \le \frac{1}{\bar{s}_\ell} \left| \frac{1}{n} \sum_i (\bar{\phi}^1_{\ell-1,i,\perp})^2 - \bar{s}^2_\ell \right|. \tag{122}$$

The empirical-product estimate Eq. (25) bounds the right side in L^{2p} by $C_{p,L} n^{-1/2}$, where the constant may depend on $\bar{s}_\ell^{-1}$ when $\bar{s}_\ell > 0$. If $\bar{s}_\ell = 0$, then $\bar{\phi}^1_{\ell-1,i,\perp} = 0$ almost surely for every i, so the second term is zero. In this case, the entire coefficient error is controlled by the preceding Jensen estimate. Thus, in both cases,

$$\left\| \mathbf{1}_{\Omega_\ell} \left(\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right) \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{123}$$

The event Ω_ℓ and the coefficient error are determined before the independent innovation z^1_ℓ is introduced. Hence

$$\left\| \mathbf{1}_{\Omega_\ell} \left(\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right) z^1_{\ell,i} \right\|_{L^{2p}} = \|z^1_{\ell,i}\|_{L^{2p}} \left\| \mathbf{1}_{\Omega_\ell} \left(\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} - \bar{s}_\ell \right) \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{124}$$

For the projection correction, conditional Gaussian moments give

$$\mathbb{E}_{z^1_\ell} \left| \frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} (P^0_{g,\ell} z^1_\ell)_i \right|^{2p} = C_p \left(\frac{\|\phi^1_{\ell-1,\perp}\|^2}{n} \right)^p \frac{|g^0_{\ell,i}|^{2p}}{\|g^0_\ell\|^{2p}}. \tag{125}$$

On Ω_ℓ, the denominator is bounded below by a positive constant times n^p. Since $\|\phi^1_{\ell-1,\perp}\| \le \|\phi^1_{\ell-1}\|$, Hölder's inequality yields

$$\mathbb{E} \left[\mathbf{1}_{\Omega_\ell} \left| \frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} (P^0_{g,\ell} z^1_\ell)_i \right|^{2p} \right] \le \frac{C_{p,L}}{n^p} \left[\mathbb{E} \left(\frac{\|\phi^1_{\ell-1}\|^2}{n} \right)^{2p} \right]^{1/2} \left[\mathbb{E} |g^0_{\ell,i}|^{4p} \right]^{1/2} \le C_{p,L} n^{-p}. \tag{126}$$

Here Jensen's inequality bounds the normalized feature moment by the average of the coordinate $4p$ moments, which are bounded by the induction hypothesis. The initial backward moments bound the other factor. Together with Eq. (123), this proves

$$\sup_i \left\| \mathbf{1}_{\Omega_\ell} \left[\frac{\|\phi^1_{\ell-1,\perp}\|}{\sqrt{n}} \big((I - P^0_{g,\ell}) z^1_\ell \big)_i - \bar{s}_\ell z^1_{\ell,i} \right] \right\|_{L^{2p}} \le C_{p,L} n^{-1/2}. \tag{127}$$

Combining the three bracket estimates in Eq. (114) proves Eq. (111) on Ω_ℓ.

We finally control the complementary event using the actual weight action. The Gaussian operator-norm estimate and the induction hypothesis give

$$\|a^1_{\ell,i}\|_{L^{4p}} \le \| \|W^0_\ell/\sqrt{n}\|_{\mathrm{op}} \|\phi^1_{\ell-1}\| \|_{L^{4p}} \le \| \|W^0_\ell/\sqrt{n}\|_{\mathrm{op}} \|_{L^{8p}} \| \|\phi^1_{\ell-1}\| \|_{L^{8p}} \le C_{p,L} \sqrt{n}. \tag{128}$$

The limiting terms have bounded $4p$ moments by Lemma E.1 and the initial moment bounds. Therefore, Hölder's inequality and Eq. (113) with $r = 4p$ give, uniformly in i,

$$\left\| \mathbf{1}_{\Omega^c_\ell} \left(a^1_{\ell,i} - \xi^1_{\ell,i} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}} \bar{g}^0_{\ell,i} \right) \right\|_{L^{2p}} \le \mathbb{P}(\Omega^c_\ell)^{1/(4p)} \left\| a^1_{\ell,i} - \xi^1_{\ell,i} - \frac{\bar{R}^{1,0}_\ell}{\sqrt{L}} \bar{g}^0_{\ell,i} \right\|_{L^{4p}} \le C_{p,L} n^{-1/2}. \tag{129}$$

This proves the unconditional estimate at every finite moment order. □

We can now close the layer induction. At each layer, state convergence gives the empirical coefficients through the empirical-product estimate Eq. (25). The preceding layer estimate then controls the reused weight action, and the updated recursion gives convergence at the next layer.

Proposition E.4 (Width convergence after one gradient step)**.** *Fix $L \geq 1$, deterministic inputs $x^0, x^1 \in \mathbb{R}^d$ with $x^0 \neq 0$, a label y, and $\eta_c \geq 0$. Suppose that $q_0 > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero with ϕ and ϕ' globally Lipschitz, and that $u \mapsto \partial_f \mathcal{L}(u, y)$ is Lipschitz continuous. Apply one gradient step with learning rate $\eta = \eta_c n$ at (x^0, y), and evaluate the updated network at x^1. The finite and limiting processes admit a joint coupling such that, for every $p \geq 2$,*

$$\max_{0 \leq \ell \leq L} \sup_i \|h_{\ell,i}^{1,n,L} - \bar{h}_{\ell,i}^{1,L}\|_{L^p} \leq C_{p,L} n^{-1/2}, \tag{130}$$

$$\max_{0 \leq \ell \leq L} \sup_i \|h_{\ell,i}^{1,n,L}\|_{L^p} \leq C_{p,L}. \tag{131}$$

In the same coupling,

$$\max_{1 \leq \ell \leq L} \sup_i \left\| a_{\ell,i}^{1,n,L} - \xi_{\ell,i}^{1,L} - \frac{\bar{R}_\ell^{1,0,L}}{\sqrt{L}} \bar{g}_{\ell,i}^{0,L} \right\|_{L^p} \leq C_{p,L} n^{-1/2}, \tag{132}$$

$$\max_{0 \leq \ell \leq L} \max_{s,t \in \{0,1\}} \|C_{\phi,\ell}^{s,t,n,L} - \bar{C}_{\phi,\ell}^{s,t,L}\|_{L^p} \leq C_{p,L} n^{-1/2}. \tag{133}$$

Proof. We suppress n, L in the process notation. Lemma E.1 gives the limiting moment bounds at fixed depth L. We prove state convergence and finite moment bounds simultaneously at every moment order by induction over layers.

At layer 0, the shared input Gaussian terms cancel, giving

$$h_{0,i}^1 - \bar{h}_{0,i}^1 = -\eta_c q_x^{1,0} \left[(\chi^0 - \bar{\chi}^0) g_{0,i}^0 + \bar{\chi}^0 (g_{0,i}^0 - \bar{g}_{0,i}^0) \right]. \tag{134}$$

Corollary D.7, Eq. (68), and Hölder's inequality at order $2p$ give the $C_p n^{-1/2}$ error. Adding this error to the limiting moment bound gives the finite moment bound at layer 0.

Suppose the assertions hold through layer $\ell - 1$ at every moment order. Lemma E.3 gives the feature covariance estimates and the corrected approximation of a_ℓ^1. For the gradient term, add and subtract intermediate products to separate the errors in χ^0, $C_{\phi,\ell-1}^{1,0}$, and $g_{\ell,i}^0$. These errors are controlled by Corollary D.7, Eq. (110), and Eq. (68), respectively. The induction hypothesis and the initial moment bounds control the remaining factors. Hölder's inequality at orders $2p$ and $3p$ therefore gives

$$\sup_i \left\| \chi^0 C_{\phi,\ell-1}^{1,0} g_{\ell,i}^0 - \bar{\chi}^0 \bar{C}_{\phi,\ell-1}^{1,0} \bar{g}_{\ell,i}^0 \right\|_{L^p} \leq C_{p,L} n^{-1/2}. \tag{135}$$

Subtracting the finite and limiting state recursions, the corrected query error and the gradient error enter with factors $L^{-1/2}$ and η_c / L, respectively. Minkowski's inequality thus gives

$$\sup_i \|h_{\ell,i}^1 - \bar{h}_{\ell,i}^1\|_{L^p} \leq \sup_i \|h_{\ell-1,i}^1 - \bar{h}_{\ell-1,i}^1\|_{L^p} + C_{p,L} n^{-1/2}.$$

The induction hypothesis proves state convergence at layer ℓ. Adding this error to the limiting moment bound gives the finite moments, closing the induction.

The corrected query estimates follow from Lemma E.3 at each layer. Finally, Lemma A.5, applied to the coupled features at every layer, gives Eq. (133). □

Corollary E.5 (Readout and output convergence)**.** *Under the assumptions and in the coupling of Proposition E.4, define*

$$\bar{v}_i^{1,L} := v_i^0 - \eta_c \bar{\chi}^0 \bar{h}_{L,i}^{0,L}, \qquad \bar{f}^{1,L} := \mathbb{E}[\bar{v}_i^{1,L} \bar{h}_{L,i}^{1,L}]. \tag{136}$$

For every $p \geq 2$,

$$\sup_i \|v_i^{1,n,L} - \bar{v}_i^{1,L}\|_{L^p} + \|f^{1,n,L} - \bar{f}^{1,L}\|_{L^p} \leq C_{p,L} n^{-1/2}. \tag{137}$$

Proof. We suppress n, L in the process notation. The readout update gives

$$v_i^1 - \bar{v}_i^1 = -\eta_c \left[(\chi^0 - \bar{\chi}^0) h_{L,i}^0 + \bar{\chi}^0 (h_{L,i}^0 - \bar{h}_{L,i}^0) \right].$$

Corollary D.7, Proposition D.2, and Hölder's inequality at order $2p$ give the readout convergence rate. The limiting readout has bounded moments, so the finite readout does as well. Applying Lemma A.5 to $(v_i^1, h_{L,i}^1)$ and the iid mean-field pairs $(\bar{v}_i^1, \bar{h}_{L,i}^1)$ then gives the output estimate. □

Remark E.6 (Scope of the constants). Write $\bar{s}_\ell^2 := \mathbb{E}[(\bar{\phi}^1_{\ell-1,i,\perp})^2]$. The width constants $C_{p,L}$ may depend on L and on $\bar{s}_\ell^{-1}$ when $\bar{s}_\ell > 0$. The case $\bar{s}_\ell = 0$ is included explicitly in Eq. (120), so no full-rank assumption on the two-feature covariance is needed. The width rate $C_{p,L}n^{-1/2}$ holds for each fixed L. Proposition E.7 below shows that the depth dependence cannot generally be removed, even when the updated mean-field moments remain uniformly bounded.

E.3 Obstruction to Depth-Uniform Width Convergence

Proposition E.7 (Failure of depth-uniform moment bounds)**.** *Consider standard Gaussian initialization, $\phi(u) = u$, the squared loss $\mathcal{L}(f, 0) = f^2/2$, and $d = 1$ with $x^0 = x^1 = 1$. Fix $\eta_c > 0$ and apply one gradient step with $\eta = \eta_c n$. For every fixed $n \geq 1$, $p \geq 2$, and $1 \leq i \leq n$,*

$$\lim_{L\to\infty} \|h^{1,n,L}_{L,i}\|_{L^p} = \infty, \qquad \sup_{L\geq 1} \|\bar{h}^{1,L}_{L,i}\|_{L^p} \leq C_p.$$

Consequently, a depth-uniform L^p width approximation is impossible under the preceding conditions.

Proof. We suppress n, L in the process notation. For the mean-field process, $\bar{f}^0 = 0$ gives $\bar{\chi}^0 = 0$. Since $x^1 = x^0$, the updated mean field equals the initial forward field at every layer: $\bar{h}^1_\ell = \bar{h}^0_\ell$. Indeed, the orthogonal feature and response correction both vanish, and the forward innovations coincide. Thus $\bar{h}^1_L$ is centered Gaussian with variance $(1 + 1/L)^L \leq e$, proving its uniform moment bound.

For the finite network, consider the event

$$\mathcal{E}_{n,L} := \{U^0_i, v^0_i, W^0_{\ell,ij} \in [1, 2] \text{ for all } i, j, \ell\}.$$

Writing $\theta := \mathbb{P}(1 \leq Z \leq 2) > 0$ for $Z \sim \mathcal{N}(0, 1)$, independence gives $\mathbb{P}(\mathcal{E}_{n,L}) = \theta^{Ln^2+2n}$. Set $\rho := 1 + \sqrt{n/L}$. On $\mathcal{E}_{n,L}$, induction through the initial forward and backward recursions gives

$$h^0_{\ell,i} \geq \rho^\ell, \qquad g^0_{\ell,i} \geq \rho^{L-\ell}, \qquad \chi^0 = f^0 \geq \rho^L.$$

The gradient update therefore implies

$$h^1_{0,i} \leq 2 - \eta_c\rho^{2L}, \qquad \frac{(W^1_\ell)_{ij}}{\sqrt{Ln}} \leq \frac{2}{\sqrt{Ln}} - \frac{\eta_c\rho^{2L-1}}{Ln}.$$

For fixed n and sufficiently large L, all coordinates of h^1_0 and all entries of $I + W^1_\ell/\sqrt{Ln}$ are negative. The coordinates of each updated state consequently have a common sign, alternating between layers, so there is no cancellation. With $m_\ell := \min_i |h^1_{\ell,i}|$, this gives

$$m_\ell \geq \left(\frac{\eta_c\rho^{2L-1}}{L} - 1 - 2\sqrt{\frac{n}{L}}\right) m_{\ell-1}, \qquad m_0 \geq \eta_c\rho^{2L} - 2.$$

For $L \geq n$, the inequality $\log(1 + t) \geq t/2$ for $0 \leq t \leq 1$ yields

$$\rho^{2L} \geq e^{\sqrt{nL}}, \qquad \rho^{2L-1} \geq \tfrac{1}{2}e^{\sqrt{nL}}.$$

Hence, for sufficiently large L,

$$|h^1_{L,i}| \geq \left(\frac{\eta_c}{4L}e^{\sqrt{nL}}\right)^L \qquad \text{on } \mathcal{E}_{n,L}.$$

Multiplying this lower bound by the event probability gives

$$\mathbb{E}|h^1_{L,i}|^p \geq \theta^{Ln^2+2n}\left(\frac{\eta_c}{4L}e^{\sqrt{nL}}\right)^{pL}. \tag{138}$$

For fixed n, the logarithm has positive leading term $p\sqrt{n}\,L^{3/2}$, while the negative terms are at most of order $L \log L$. It therefore tends to infinity, proving the finite-network moment divergence.

Finally, under any coupling, the reverse triangle inequality gives

$$\|h^1_{L,i} - \bar{h}^1_{L,i}\|_{L^p} \geq \|h^1_{L,i}\|_{L^p} - \|\bar{h}^1_{L,i}\|_{L^p} \longrightarrow \infty.$$

Thus the width approximation cannot have a constant independent of depth. □

Remark E.8 (Joint width–depth scaling). In Eq. (138), the logarithmic amplification has order $\sqrt{n}\,L^{3/2}$, while the probability cost has order n^2L. The same argument therefore gives moment divergence along sequences with $L/n^3 \to \infty$. This is a sufficient regime for divergence, not a sharp threshold: the bound may not establish stability when $L = o(n^3)$.

E.4 Response Closure for the Updated Mean Field

We study the depth limit of the updated mean field in Eq. (106). Its recursion involves both the feature kernels $\bar{C}^{1,0,L}_{\phi,\ell-1}$ and the response $\bar{R}^{1,0,L}_{\ell}$. We first express this response through a derivative kernel and an accumulated deterministic coefficient, obtaining closed dynamics suitable for the depth limit.

Throughout the analysis, we set

$$\Delta = L^{-1}, \qquad t_\ell = \ell\Delta, \qquad \lambda = \eta_c \partial_f \mathcal{L}(0, y), \qquad q_x^{1,0} = \langle x^1, x^0\rangle/d, \qquad q_x^{1,1} = \|x^1\|^2/d.$$

All expectations refer to one mean-field coordinate. Recall the updated recursion and response:

$$\begin{aligned}
\bar{h}_0^{1,L} &= \xi_0^1 - \lambda q_x^{1,0} \bar{g}_0^{0,L}, \\
\bar{h}_\ell^{1,L} &= \bar{h}_{\ell-1}^{1,L} + \sqrt{\Delta}\, \xi_\ell^{1,L} + \Delta\big(\bar{R}_\ell^{1,0,L} - \lambda \bar{C}^{1,0,L}_{\phi,\ell-1}\big)\bar{g}_\ell^{0,L}, \\
\bar{R}_\ell^{1,0,L} &= \frac{\mathbb{E}[\zeta_\ell^{0,L}\phi(\bar{h}_{\ell-1}^{1,L})]}{\sqrt{\Delta}\, r_\ell^L}.
\end{aligned} \tag{139}$$

Here $\bar{h}^{0,L}, \bar{g}^{0,L}$ and r^L are the initialization mean fields and backward variance.

Lemma E.9 (Exact finite-depth response closure)**.** *Suppose $q_0 = \|x^0\|^2/d > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero, and that ϕ and ϕ' are globally Lipschitz. Fix $\eta_c \geq 0$ and a label y with finite $\partial_f \mathcal{L}(0, y)$. Define the accumulated backward coefficient*

$$\alpha_\ell^L := -\lambda q_x^{1,0} + \Delta \sum_{j=1}^{\ell} \big(\bar{R}_j^{1,0,L} - \lambda \bar{C}^{1,0,L}_{\phi,j-1}\big), \qquad 0 \leq \ell \leq L.$$

For $1 \leq \ell \leq L$,

$$\bar{R}_\ell^{1,0,L} = \alpha_{\ell-1}^L \mathbb{E}[\phi'(\bar{h}_{\ell-1}^{1,L})\phi'(\bar{h}_{\ell-1}^{0,L})]. \tag{140}$$

Consequently, $\alpha_0^L = -\lambda q_x^{1,0}$ and

$$\alpha_\ell^L = \alpha_{\ell-1}^L + \Delta\left(\alpha_{\ell-1}^L \mathbb{E}[\phi'(\bar{h}_{\ell-1}^{1,L})\phi'(\bar{h}_{\ell-1}^{0,L})] - \lambda \bar{C}^{1,0,L}_{\phi,\ell-1}\right). \tag{141}$$

Proof. Fix $1 \leq \ell \leq L$. For $0 \leq j < \ell$, the backward recursion gives

$$\frac{\partial \bar{g}_j^{0,L}}{\partial \zeta_\ell^{0,L}} = \sqrt{\Delta}\, \phi'(\bar{h}_{\ell-1}^{0,L}).$$

Since the forward innovations $\xi_j^{1,L}$ are independent of $\zeta_\ell^{0,L}$, differentiating the input update and the first $\ell - 1$ updated increments, with all expectation coefficients held fixed, gives

$$\frac{\partial \bar{h}_{\ell-1}^{1,L}}{\partial \zeta_\ell^{0,L}} = \sqrt{\Delta}\left[-\lambda q_x^{1,0} + \Delta \sum_{j=1}^{\ell-1} \big(\bar{R}_j^{1,0,L} - \lambda \bar{C}^{1,0,L}_{\phi,j-1}\big)\right] \phi'(\bar{h}_{\ell-1}^{0,L}) = \sqrt{\Delta}\, \alpha_{\ell-1}^L \phi'(\bar{h}_{\ell-1}^{0,L}).$$

Since $\zeta_\ell^{0,L}$ is centered Gaussian with variance r_ℓ^L, Gaussian integration by parts, conditional on all other Gaussian inputs and followed by expectation, gives

$$\mathbb{E}[\zeta_\ell^{0,L}\phi(\bar{h}_{\ell-1}^{1,L})] = r_\ell^L \sqrt{\Delta}\, \alpha_{\ell-1}^L \mathbb{E}[\phi'(\bar{h}_{\ell-1}^{1,L})\phi'(\bar{h}_{\ell-1}^{0,L})].$$

Boundedness of ϕ' and the fixed-depth moment bounds justify this step. Dividing by $\sqrt{\Delta}\, r_\ell^L$ proves Eq. (140). Substituting this identity into the definition of α^L gives Eq. (141). □

E.5 Diffusion Coefficients and Uniform Discrete Bounds

The response closure identifies the drift. We next control the diffusion coefficients arising from the Gaussian covariance decomposition. The same regression rule will be used for the discrete and continuous mean fields. For random variables $F_0, F_1 \in L^2$ with $\|F_0\|_{L^2} > 0$, define

$$\sigma_0(F_0, F_1) := \frac{\mathbb{E}[F_0 F_1]}{\|F_0\|_{L^2}}, \qquad \sigma_1(F_0, F_1) := \left\| F_1 - \frac{\mathbb{E}[F_0 F_1]}{\mathbb{E}[F_0^2]} F_0 \right\|_{L^2}. \tag{142}$$

In particular, $\sigma_0(F_0, F_1)^2 + \sigma_1(F_0, F_1)^2 = \mathbb{E}[F_1^2]$. The following lemma supplies the regularity needed for well-posedness and the later depth comparison.

Lemma E.10 (Lipschitz continuity of the diffusion coefficients)**.** *Let $F_0, F_1, \widetilde{F}_0, \widetilde{F}_1 \in L^2$ satisfy*

$$\|F_0\|_{L^2}, \|\widetilde{F}_0\|_{L^2} \geq \rho > 0, \qquad \|F_1\|_{L^2}, \|\widetilde{F}_1\|_{L^2} \leq R.$$

For $j = 0, 1$,

$$|\sigma_j(F_0, F_1) - \sigma_j(\widetilde{F}_0, \widetilde{F}_1)| \leq C_{\rho,R}\big(\|F_0 - \widetilde{F}_0\|_{L^2} + \|F_1 - \widetilde{F}_1\|_{L^2}\big).$$

No positive lower bound on σ_1 is required.

Proof. Set $u = F_0/\|F_0\|_{L^2}$ and $\widetilde{u} = \widetilde{F}_0/\|\widetilde{F}_0\|_{L^2}$. The triangle and reverse triangle inequalities give

$$\|u - \widetilde{u}\|_{L^2} \leq 2\rho^{-1}\|F_0 - \widetilde{F}_0\|_{L^2}. \tag{143}$$

Let $P_u f = \langle f, u\rangle_{L^2} u$ be the orthogonal projection onto the span of u. Then

$$\sigma_0(F_0, F_1) = \langle F_1, u\rangle_{L^2}, \qquad \sigma_1(F_0, F_1) = \|(I - P_u)F_1\|_{L^2}.$$

Adding and subtracting $\langle f, u\rangle_{L^2}\widetilde{u}$ shows that $\|P_u - P_{\widetilde{u}}\|_{\mathrm{op}} \leq 2\|u - \widetilde{u}\|_{L^2}$. Cauchy–Schwarz and the reverse triangle inequality therefore give

$$|\sigma_0(F_0, F_1) - \sigma_0(\widetilde{F}_0, \widetilde{F}_1)| \leq \|F_1 - \widetilde{F}_1\|_{L^2} + R\|u - \widetilde{u}\|_{L^2},$$
$$|\sigma_1(F_0, F_1) - \sigma_1(\widetilde{F}_0, \widetilde{F}_1)| \leq \|F_1 - \widetilde{F}_1\|_{L^2} + 2R\|u - \widetilde{u}\|_{L^2}.$$

Substituting Eq. (143) into these two inequalities proves the claimed bounds. □

We now represent the closed discrete dynamics using Brownian increments, as at initialization. Introduce an independent Brownian motion B^1, preserving the covariance of the updated and initial forward innovations. Write $\Delta B^j_\ell = B^j_{t_\ell} - B^j_{t_{\ell-1}}$ for $j \in \{0, 1\}$ and $\Delta\widetilde{B}_\ell = \widetilde{B}_{t_\ell} - \widetilde{B}_{t_{\ell-1}}$. Then

$$\xi^{1,L}_\ell = \sum_{j=0}^{1} \sigma_j\big(\phi(\bar{h}^{0,L}_{\ell-1}), \phi(\bar{h}^{1,L}_{\ell-1})\big)\Delta B^j_\ell/\sqrt{\Delta}, \qquad \zeta^{0,L}_\ell = \sqrt{r^L_\ell}\,\Delta\widetilde{B}_\ell/\sqrt{\Delta}. \tag{144}$$

By the response closure, the updated state satisfies

$$\bar{h}^{1,L}_0 = \xi^1_0 + \alpha^L_0\bar{g}^{0,L}_0, \qquad \bar{h}^{1,L}_\ell = \bar{h}^{1,L}_{\ell-1} + \sqrt{\Delta}\,\xi^{1,L}_\ell + (\alpha^L_\ell - \alpha^L_{\ell-1})\bar{g}^{0,L}_\ell, \tag{145}$$

with α^L given by Eq. (141). All expectation coefficients are deterministic. The Gaussian innovations are independent across layers. However, the updated state $\bar{h}^{1,L}_{\ell-1}$ may depend on earlier innovation $\zeta^{0,L}_\ell$ through the backward fields $\bar{g}^{0,L}_j$ at layers $j < \ell$.

Before defining the continuous limit, we establish uniform bounds for these discrete mean fields.

Lemma E.11 (Uniform finite-depth bounds)**.** *Suppose $q_0 = \|x^0\|^2/d > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero, and that ϕ and ϕ' are globally Lipschitz. Fix $\eta_c \geq 0$ and a label y with finite $\partial_f\mathcal{L}(0, y)$. For every $p \geq 2$,*

$$\max_{0\leq\ell\leq L}|\alpha^L_\ell| + \max_{1\leq\ell\leq L}\left|\frac{\alpha^L_\ell - \alpha^L_{\ell-1}}{\Delta}\right| + \left\|\max_{0\leq\ell\leq L}|\bar{h}^{1,L}_\ell|\right\|_{L^p} \leq C_p, \tag{146}$$

where C_p is independent of L.

Proof. The projection decomposition gives $\sigma_0(F_0, F_1)^2 + \sigma_1(F_0, F_1)^2 = \mathbb{E}[F_1^2]$. Since ϕ has linear growth, ϕ' is bounded, and Lemma D.1 bounds the initial feature moments uniformly, Eq. (141) therefore gives

$$\sum_{j=0}^{1}\sigma_j\big(\phi(\bar{h}^{0,L}_{\ell-1}), \phi(\bar{h}^{1,L}_{\ell-1})\big)^2 + \left|\frac{\alpha^L_\ell - \alpha^L_{\ell-1}}{\Delta}\right|^2 \leq C\big(1 + |\alpha^L_{\ell-1}|^2 + \mathbb{E}|\bar{h}^{1,L}_{\ell-1}|^2\big). \tag{147}$$

Iterating Eq. (145) gives

$$\bar{h}^{1,L}_\ell = \xi^1_0 + \alpha^L_0\bar{g}^{0,L}_0 + \sqrt{\Delta}\sum_{j=1}^{\ell}\xi^{1,L}_j + \sum_{j=1}^{\ell}(\alpha^L_j - \alpha^L_{j-1})\bar{g}^{0,L}_j.$$

The Gaussian partial sums are Brownian integrals with deterministic coefficients. Lemma A.9 gives

$$\mathbb{E}\left[\max_{0\le k\le \ell}\left|\sqrt{\Delta}\sum_{j=1}^{k}\xi_j^{1,L}\right|^2\right] \le C\Delta\sum_{j=1}^{\ell}\bar{C}_{\phi,j-1}^{1,1,L}.$$

For the drift sums, Minkowski's, the initial backward moment bounds, and Cauchy–Schwarz give

$$\left\|\max_{0\le k\le \ell}\left|\sum_{j=1}^{k}(\alpha_j^L-\alpha_{j-1}^L)\bar{g}_j^{0,L}\right|\right\|_{L^2}^2 \le \left(\sum_{j=1}^{\ell}|\alpha_j^L-\alpha_{j-1}^L|\|\bar{g}_j^{0,L}\|_{L^2}\right)^2 \le \frac{C}{\Delta}\sum_{j=1}^{\ell}|\alpha_j^L-\alpha_{j-1}^L|^2.$$

Define

$$\mathcal{E}_\ell^L := \max_{0\le k\le \ell}|\alpha_k^L|^2 + \mathbb{E}\left[\max_{0\le k\le \ell}|\bar{h}_k^{1,L}|^2\right].$$

Together with Eq. (147), the initial moment bounds, and the recursion for α^L, these estimates imply

$$\mathcal{E}_\ell^L \le C + C\Delta\sum_{j=0}^{\ell-1}\mathcal{E}_j^L.$$

Discrete Gronwall bounds $\mathcal{E}_\ell^L$ uniformly in L. The growth estimate then bounds the normalized increments $(\alpha_\ell^L-\alpha_{\ell-1}^L)/\Delta$ and the deterministic diffusion coefficients uniformly. Applying Lemma A.9 in L^p to the Gaussian sums and Minkowski's inequality to the drift sums, using the initial backward L^p moment bounds, proves Eq. (146). □

E.6 The Continuous-Depth Integral System

The closed Brownian recursion now suggests a continuous integral system. Retain the initialization fields H^0, G^0 and variances q, r. For a candidate updated field H^1, define

$$\begin{aligned} C_\phi^{1,0}(t) &:= \mathbb{E}[\phi(H_t^1)\phi(H_t^0)], & C_\phi^{1,1}(t) &:= \mathbb{E}[\phi(H_t^1)^2],\\ C_{\phi'}^{1,0}(t) &:= \mathbb{E}[\phi'(H_t^1)\phi'(H_t^0)], & \sigma_j(t) &:= \sigma_j\big(\phi(H_t^0),\phi(H_t^1)\big), \quad j=0,1. \end{aligned} \tag{148}$$

Replacing the discrete sums by integrals gives the candidate system

$$\alpha(t) = -\lambda q_x^{1,0} + \int_0^t\left[\alpha(s)C_{\phi'}^{1,0}(s) - \lambda C_\phi^{1,0}(s)\right]ds, \tag{149}$$

$$H_t^1 = \xi_0^1 - \lambda q_x^{1,0}G_0^0 + \sum_{j=0}^{1}\int_0^t \sigma_j(s)\,dB_s^j + \int_0^t\left[\alpha(s)C_{\phi'}^{1,0}(s) - \lambda C_\phi^{1,0}(s)\right]G_s^0\,ds, \tag{150}$$

The coefficients are deterministic. For any continuous solution, the first equation yields $\alpha \in C^1$ and

$$\alpha'(t) = \alpha(t)C_{\phi'}^{1,0}(t) - \lambda C_\phi^{1,0}(t). \tag{151}$$

Thus the field drift can also be written as $\int_0^t \alpha'(s)G_s^0\,ds$. The stochastic integrals are ordinary Itô integrals, and the drift is a pathwise Lebesgue integral. The updated field need not be adapted, because its initial value contains G_0^0.

A strong solution is a deterministic continuous α and a continuous H^1, measurable in the prescribed Gaussian primitives, satisfying these equations with finite L^2 norm of the path supremum on each interval of existence. Pathwise uniqueness refers to solutions with the same primitives. A weak solution satisfies the equations on any space carrying the prescribed primitive laws, without the measurability requirement.

Proposition E.12 (Global strong existence and uniqueness)**.** *Suppose $q_0 = \|x^0\|^2/d > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero, and that ϕ and ϕ' are globally Lipschitz. Fix $\eta_c \ge 0$ and a label y with finite $\partial_f\mathcal{L}(0,y)$. The system Eq. (149)–Eq. (150) has a unique global strong solution (α, H^1) on $[0,1]$, with deterministic $\alpha \in C^1([0,1])$. For every $p \ge 2$,*

$$\|\alpha\|_\infty + \|\alpha'\|_\infty + \left\|\sup_{t\le 1}|H_t^1|\right\|_{L^p} \le C_p. \tag{152}$$

A weak solution also exists, and its joint law with the initial fields and prescribed primitives is unique.

Proof. We first construct a local solution by Picard iteration. Write $\alpha_0 = -\lambda q_x^{1,0}$ and $H_0^1 = \xi_0^1 + \alpha_0 G_0^0$. On $C([0,T]) \times L^2(\Omega; C([0,T]))$, use the norm

$$\|(\alpha, H^1)\|_T := \|\alpha\|_{\infty,[0,T]} + \left\| \sup_{t \le T} |H_t^1| \right\|_{L^2}.$$

Let $\mathcal{T}$ be the map defined by the right sides of Eq. (149) and Eq. (150). Consider the closed radius-one ball around the constant initial paths, with the prescribed initial values. Every candidate satisfies

$$\|(\alpha, H^1)\|_T \le 1 + |\alpha_0| + \|H_0^1\|_{L^2} =: R.$$

For two candidates (α, H^1) and $(\widetilde{\alpha}, \widetilde{H}^1)$, the initial feature nondegeneracy from Lemma D.1 and Lemma E.10 control the diffusion coefficients. Cauchy–Schwarz and the Lipschitz assumptions give

$$\begin{aligned} |C_\phi^{1,0}(t) - \widetilde{C}_\phi^{1,0}(t)| &\le K_0 \|\phi(H_t^0)\|_{L^2} \|H_t^1 - \widetilde{H}_t^1\|_{L^2}, \\ |C_{\phi'}^{1,0}(t) - \widetilde{C}_{\phi'}^{1,0}(t)| &\le K_0 K_1 \|H_t^1 - \widetilde{H}_t^1\|_{L^2}. \end{aligned}$$

Combining these estimates with the bounded candidate norms controls the drift and diffusion differences:

$$\begin{aligned} &\big|\alpha(t) C_{\phi'}^{1,0}(t) - \widetilde{\alpha}(t) \widetilde{C}_{\phi'}^{1,0}(t)\big| \\ &\quad + |\lambda|\, |C_\phi^{1,0}(t) - \widetilde{C}_\phi^{1,0}(t)| + \sum_{j=0}^{1} |\sigma_j(t) - \widetilde{\sigma}_j(t)| \\ &\quad \le C_R\big(|\alpha(t) - \widetilde{\alpha}(t)| + \|H_t^1 - \widetilde{H}_t^1\|_{L^2}\big). \end{aligned} \tag{153}$$

Lemma A.9 applies to the deterministic stochastic integrands. Minkowski's inequality handles the drift using $\sup_t \|G_t^0\|_{L^2} \le C$. Consequently,

$$\|\mathcal{T}(\alpha, H^1) - \mathcal{T}(\widetilde{\alpha}, \widetilde{H}^1)\|_T \le C_R(\sqrt{T} + T) \|(\alpha, H^1) - (\widetilde{\alpha}, \widetilde{H}^1)\|_T.$$

The bounded coefficients on the ball give the corresponding invariance estimate

$$\|\mathcal{T}(\alpha, H^1) - (\alpha_0, H_0^1)\|_T \le C_R(\sqrt{T} + T).$$

For sufficiently small T, the map preserves the ball and is a contraction. The Banach fixed-point theorem gives local existence, and the same difference estimate gives local pathwise uniqueness.

To continue the solution, we bound the product norm directly. As in Eq. (147), the identity for the Gaussian coefficients and the growth bounds on ϕ, ϕ' give

$$\sum_{j=0}^{1} \sigma_j(t)^2 + |\alpha'(t)|^2 \le C\big(1 + |\alpha(t)|^2 + \mathbb{E}|H_t^1|^2\big). \tag{154}$$

Indeed, the first sum is $C_\phi^{1,1}(t)$, while $|C_{\phi'}^{1,0}| \le K_0^2$ and $|C_\phi^{1,0}| \le C\sqrt{1 + \mathbb{E}|H_t^1|^2}$ bound α'. Since α' is deterministic, for $t \le 1$,

$$\left\| \sup_{u \le t} \left| \int_0^u \alpha'(s) G_s^0 \, ds \right| \right\|_{L^2}^2 \le \left(\int_0^t |\alpha'(s)| \|G_s^0\|_{L^2} \, ds \right)^2 \le C \int_0^t |\alpha'(s)|^2 \, ds.$$

Thus, with

$$\mathcal{E}(t) := \sup_{u \le t} |\alpha(u)|^2 + \mathbb{E} \sup_{u \le t} |H_u^1|^2,$$

Lemma A.9, the initial moment bounds, and Eq. (154) give

$$\mathcal{E}(t) \le C + C \int_0^t \mathcal{E}(s) \, ds.$$

Gronwall bounds $\mathcal{E}$ independently of the local existence interval. In particular, for every such $T \le 1$,

$$\|\alpha\|_{\infty,[0,T]} + \left\| \sup_{t \le T} |H_t^1| \right\|_{L^2} \le C.$$

The local construction can therefore be restarted on successive intervals of a common positive length, extending the solution to $[0,1]$.

The growth estimate now bounds α' and the diffusion coefficients on $[0,1]$. Applying Lemma A.9 in L^p and using the initial backward L^p moment bounds in Eq. (150) proves Eq. (152). Local uniqueness on successive intervals gives global pathwise uniqueness. Picard iteration constructs a measurable functional of the primitives; the same construction on any weak-solution space identifies every solution with that functional. This proves weak existence and uniqueness of the joint law. □

Both the discrete approximation and the continuous limit now have uniform moment bounds, and the limiting system is well posed. We next compare them under the common Brownian construction to obtain the depth-convergence rate.

E.7 Depth Convergence of the Updated Mean Field

Use the same inputs and Brownian motions for the discrete and continuous systems, together with the initialization stochastic interpolations $\bar{h}^{0,L}, \bar{g}^{0,L}$. For $s \in (t_{\ell-1}, t_\ell]$, freeze the diffusion coefficients at the left grid point:

$$\sigma_j^L(s) := \sigma_j\big(\phi(\bar{h}_{\ell-1}^{0,L}), \phi(\bar{h}_{\ell-1}^{1,L})\big), \qquad j = 0, 1. \tag{155}$$

Let α^L be the piecewise affine interpolation:

$$\alpha^L(t) = \alpha_{\ell-1}^L + \frac{t - t_{\ell-1}}{\Delta}(\alpha_\ell^L - \alpha_{\ell-1}^L), \qquad t \in [t_{\ell-1}, t_\ell]. \tag{156}$$

By Eq. (141), its derivative satisfies

$$(\alpha^L)'(s) = \frac{\alpha_\ell^L - \alpha_{\ell-1}^L}{\Delta} = \alpha_{\ell-1}^L \mathbb{E}[\phi'(\bar{h}_{\ell-1}^{1,L})\phi'(\bar{h}_{\ell-1}^{0,L})] - \lambda \bar{C}_{\phi,\ell-1}^{1,0,L} \tag{157}$$

for almost every $s \in (t_{\ell-1}, t_\ell]$.

On each interval $[t_{\ell-1}, t_\ell]$, define the stochastic interpolation by

$$\bar{h}^{1,L}(t) = \bar{h}_{\ell-1}^{1,L} + \sum_{j=0}^{1} \int_{t_{\ell-1}}^{t} \sigma_j^L(s)\, dB_s^j + \big(\alpha^L(t) - \alpha_{\ell-1}^L\big)\bar{g}_\ell^{0,L}, \tag{158}$$

starting from $\bar{h}_0^{1,L} = \xi_0^1 - \lambda q_x^{1,0} \bar{g}_0^{0,L}$. At $t = t_\ell$, these formulas recover $\alpha_\ell^L, \bar{h}_\ell^{1,L}$. The diffusion coefficients and $(\alpha^L)'$ use layer $\ell - 1$, while the backward factor uses layer ℓ, as required by the discrete recursion.

Proposition E.13 (Strong depth convergence after one update)**.** *Suppose $q_0 = \|x^0\|^2/d > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero, and that ϕ and ϕ' are globally Lipschitz. Fix $\eta_c \geq 0$ and a label y with finite $\partial_f \mathcal{L}(0, y)$. For every $p \geq 2$, the common Brownian coupling satisfies*

$$\|\alpha^L - \alpha\|_\infty + \Big\| \sup_{t \leq 1} |\bar{h}^{1,L}(t) - H_t^1| \Big\|_{L^p} \leq C_p L^{-1/2}. \tag{159}$$

In particular, this bound holds for the maximum error over the depth grid. In the supremum path metric,

$$W_p\big(\mathrm{Law}(\bar{h}^{0,L}, \bar{g}^{0,L}, \bar{h}^{1,L}), \mathrm{Law}(H^0, G^0, H^1)\big) \leq C_p L^{-1/2}. \tag{160}$$

Proof. We first prove convergence in L^2. Define

$$e(t) := \sup_{u \leq t} |\alpha^L(u) - \alpha(u)|^2 + \mathbb{E} \sup_{u \leq t} |\bar{h}^{1,L}(u) - H_u^1|^2.$$

The bounds from Proposition E.12, together with Lemma A.9 and Minkowski's inequality, give

$$|\alpha(t) - \alpha(s)| \leq C|t - s|, \qquad \|H_t^1 - H_s^1\|_{L^p} \leq C_p |t - s|^{1/2}.$$

For $s \in [t_{\ell-1}, t_\ell]$, adding and subtracting the continuous fields at $t_{\ell-1}$ therefore gives

$$|\alpha_{\ell-1}^L - \alpha(s)| \leq \sqrt{e(s)} + C\Delta, \qquad \|\bar{h}_{\ell-1}^{1,L} - H_s^1\|_{L^2} \leq \sqrt{e(s)} + C\sqrt{\Delta}.$$

The initialization forward depth estimate Eq. (90) and increment bound Eq. (85) similarly control $\|\bar{h}^{0,L}_{\ell-1} - H^0_s\|_{L^2}$ by $C\sqrt{\Delta}$. Since the initialization feature variance is uniformly positive and the updated fields have uniformly bounded moments, Lemma E.10 controls the diffusion differences. The kernel estimates in Eq. (153), together with Eq. (157), control the derivative difference. Squaring and adding gives, for almost every s,

$$\sum_{j=0}^{1} |\sigma^L_j(s) - \sigma_j(s)|^2 + |(\alpha^L)'(s) - \alpha'(s)|^2 \le C(e(s) + \Delta). \tag{161}$$

For $t \in [t_{\ell-1}, t_\ell]$, summing the layer increments in Eq. (158) and subtracting Eq. (150) gives

$$\begin{aligned}
\bar{h}^{1,L}(t) - H^1_t = & -\lambda q^{1,0}_x(\bar{g}^{0,L}_0 - G^0_0) \\
& + \sum_{j=0}^{1} \int_0^t [\sigma^L_j(s) - \sigma_j(s)]\, dB^j_s \\
& + \sum_{k=1}^{\ell-1} \int_{t_{k-1}}^{t_k} \left[(\alpha^L)'(s)\bar{g}^{0,L}_k - \alpha'(s)G^0_s\right] ds \\
& + \int_{t_{\ell-1}}^{t} \left[(\alpha^L)'(s)\bar{g}^{0,L}_\ell - \alpha'(s)G^0_s\right] ds.
\end{aligned}$$

The initial term has squared L^2 norm at most $C\Delta$ by Eq. (91). For the Brownian integral, Lemma A.9 and Eq. (161) give

$$\mathbb{E} \sup_{u \le t} \left| \sum_{j=0}^{1} \int_0^u [\sigma^L_j(s) - \sigma_j(s)]\, dB^j_s \right|^2 \le C \int_0^t (e(s) + \Delta)\, ds.$$

For the drift, the initialization estimates Eq. (91) and Eq. (85) bound $\|\bar{g}^{0,L}_\ell - G^0_s\|_{L^2}$ by $C\sqrt{\Delta}$ on $(t_{\ell-1}, t_\ell]$. Adding and subtracting $(\alpha^L)'(s)G^0_s$, and using Eq. (161), the preceding initialization bounds, and the uniform backward moment and derivative bounds, gives

$$\|(\alpha^L)'(s)\bar{g}^{0,L}_\ell - \alpha'(s)G^0_s\|_{L^2} \le C\sqrt{e(s) + \Delta}.$$

Applying Minkowski's and Cauchy–Schwarz inequalities over the completed intervals and the current partial interval bounds the squared maximal drift error by $C\int_0^t (e(s) + \Delta)\, ds$. Integrating the derivative difference in Eq. (161) gives the same bound for $\sup_{u \le t} |\alpha^L(u) - \alpha(u)|^2$, since $\alpha^L(0) = \alpha(0)$. Combining these estimates yields

$$e(t) \le C\Delta + C \int_0^t e(s)\, ds.$$

Gronwall therefore gives $\sup_{t \le 1} e(t) \le C\Delta$.

The coefficient differences in Eq. (161) are deterministic, so their essential suprema are now bounded by $C\sqrt{\Delta}$. Repeating the Brownian and drift estimates in L^p, using the initialization L^p depth and increment bounds, proves Eq. (159). The grid estimate follows by restriction. Together with the initialization path estimates, the common coupling also proves Eq. (160). □

Remark E.14 (Interpolation and adaptedness). The uniform path rate concerns the stochastic interpolation above and hence also the grid values. It is not asserted for ordinary piecewise affine interpolation, which omits the within-layer Brownian fluctuations. Throughout the proof, BDG is applied only to Wiener integrals with deterministic coefficients. The backward factors are handled in pathwise Lebesgue integrals; no adaptedness of the updated field is required.

E.8 Kernels, Readout, and the Iterated Limit

Corollary E.15 (Kernel, response, and output convergence)**.** *Under the assumptions and coupling of Proposition E.13, use $C_\phi^{s,t}(u) := \mathbb{E}[\phi(H_u^s)\phi(H_u^t)]$ for $s,t \in \{0,1\}$ and define*

$$\bar{C}_{\phi',\ell}^{1,0,L} := \mathbb{E}[\phi'(\bar{h}_\ell^{1,L})\phi'(\bar{h}_\ell^{0,L})], \qquad R^{1,0}(t) := \alpha(t)C_{\phi'}^{1,0}(t).$$

Then

$$\max_{0\le\ell\le L}\left(\max_{s,t\in\{0,1\}}|\bar{C}_{\phi,\ell}^{s,t,L} - C_\phi^{s,t}(t_\ell)| + |\bar{C}_{\phi',\ell}^{1,0,L} - C_{\phi'}^{1,0}(t_\ell)|\right) \le CL^{-1/2}, \tag{162}$$

$$\max_{1\le\ell\le L}|\bar{R}_\ell^{1,0,L} - R^{1,0}(t_{\ell-1})| \le CL^{-1/2}. \tag{163}$$

For $V^1 := v^0 - \lambda H_1^0$ and $\bar{v}^{1,L} := v^0 - \lambda\bar{h}_L^{0,L}$, set $f_\infty^1 := \mathbb{E}[V^1H_1^1]$ and $\bar{f}^{1,L} := \mathbb{E}[\bar{v}^{1,L}\bar{h}_L^{1,L}]$. For every $p \ge 2$,

$$\|\bar{v}^{1,L} - V^1\|_{L^p} \le C_pL^{-1}, \qquad |\bar{f}^{1,L} - f_\infty^1| \le CL^{-1/2}. \tag{164}$$

The outputs also satisfy

$$f_\infty^1 = \alpha(1) - \lambda\left(q_x^{1,0} + \int_0^1 C_\phi^{1,0}(s)\,ds\right), \tag{165}$$

$$\bar{f}^{1,L} = \alpha_L^L - \lambda\left(q_x^{1,0} + \Delta\sum_{j=1}^{L}\bar{C}_{\phi,j-1}^{1,0,L}\right).$$

The initial mean-field output is zero at every depth and in the continuous-depth limit.

Proof. For coupled square-integrable variables,

$$|\mathbb{E}[X_LY_L] - \mathbb{E}[XY]| \le \|X_L - X\|_{L^2}\|Y_L\|_{L^2} + \|X\|_{L^2}\|Y_L - Y\|_{L^2}.$$

Apply this inequality to the features, using their uniform moments, Lipschitz continuity of ϕ, and the strong depth estimates. The derivative kernels follow in the same way from boundedness and Lipschitz continuity of ϕ'. The exact identity $\bar{R}_\ell^{1,0,L} = \alpha_{\ell-1}^L\bar{C}_{\phi',\ell-1}^{1,0,L}$ then gives the response rate. The readout rate is the initial forward rate multiplied by $|\lambda|$. The product estimate above, applied to the readout and final updated state, gives the output rate.

For the scalar formulas, the backward representation gives $\mathbb{E}[v^0G_s^0] = 1$ and $\mathbb{E}[H_t^0G_s^0] = 0$ for every s,t. The first identity uses independence of v^0 and the backward Brownian integral; the second uses its conditional centering given the forward primitives. Taking the corresponding cross moments in Eq. (150) yields

$$\mathbb{E}[v^0H_t^1] = \alpha(t), \qquad \mathbb{E}[H_t^0H_t^1] = q_x^{1,0} + \int_0^t\sqrt{C_\phi^{0,0}(s)}\sigma_0(s)\,ds = q_x^{1,0} + \int_0^t C_\phi^{1,0}(s)\,ds.$$

The same calculation in the discrete recursion gives $\mathbb{E}[v^0\bar{h}_\ell^{1,L}] = \alpha_\ell^L$ and the corresponding cross moment with the sum in Eq. (165). Substitution in the output definitions proves the formulas. The initial outputs vanish by independence and centering of v^0. □

Corollary E.16 (Iterated width–depth limit)**.** *Suppose $q_0 > 0$, that $\phi \in C^1(\mathbb{R})$ is not identically zero, that ϕ, ϕ' are globally Lipschitz, and that $u \mapsto \partial_f\mathcal{L}(u,y)$ is globally Lipschitz. Fix $\eta_c \ge 0$ and use learning rate $\eta = \eta_c n$. The width and depth couplings can be combined so that, for every $p \ge 2$, uniformly in $0 \le \ell \le L$ and the coordinate i,*

$$\|h_{\ell,i}^{0,n,L} - H_{t_\ell,i}^0\|_{L^p} + \|g_{\ell,i}^{0,n,L} - G_{t_\ell,i}^0\|_{L^p} + \|h_{\ell,i}^{1,n,L} - H_{t_\ell,i}^1\|_{L^p} \le C_{p,L}n^{-1/2} + C_pL^{-1/2}, \tag{166}$$

$$\|f^{1,n,L} - f_\infty^1\|_{L^p} \le C_{p,L}n^{-1/2} + CL^{-1/2}.$$

Here (H_i^0, G_i^0, H_i^1) are independent coordinate copies of the limiting system. Taking width first and then depth gives convergence at every fixed finite collection of coordinates and depths, and convergence of the output to f_∞^1. A simultaneous limit also follows along sequences with $L(n) \to \infty$ and $C_{p,L(n)}n^{-1/2} \to 0$ for the required moment orders.

Proof. At fixed L, the width coupling and the Brownian depth coupling have the same marginal law on the full discrete mean-field skeleton $(\bar{h}^{0,L}_{\ell,i}, \bar{g}^{0,L}_{\ell,i}, \bar{h}^{1,L}_{\ell,i}, v^0_i)_{0\le\ell\le L,\,1\le i\le n}$. Gluing the couplings along this common marginal preserves both approximations. The triangle inequality gives Eq. (166) and the output bound. For depths off the grid, choose nearest grid points and use the limiting L^p increment bounds. Applying these estimates to a fixed finite collection of coordinates and depths proves the iterated limit and the stated simultaneous limit. □

Remark E.17 (Scope of the depth estimates). The uniform bounds above concern the infinite-width mean fields. They do not remove the depth dependence of the finite-network constants $C_{p,L}$. Proposition E.7 shows that finite-network updated moments can diverge with depth even when the mean-field moments stay bounded. The sequential limit is therefore essential to the unconditional statement; the simultaneous-limit condition above is not an explicit growth rule for $L(n)$.

F Mean-Field Dynamics over Multiple SGD Updates

We consider a fixed finite number k of SGD updates with learning rate $\eta = \eta_c n$. For a fixed sample sequence, we first describe the finite-width dynamics, their fixed-depth mean-field counterpart, and the candidate continuous-depth equations. The next two subsections address width convergence and depth convergence under suitable assumptions. Response closure and the interpretation of the continuous backward stochastic term are deferred to the depth analysis.

We consider a fixed finite number k of SGD updates with learning rate $\eta = \eta_c n$. For a fixed sample sequence, we first describe the finite-width dynamics, their fixed-depth mean-field counterpart, and the candidate continuous-depth equations.

F.1 Discrete and Continuous Mean-Field Dynamics

For $0 \le j \le k$, write

$$\phi^j_\ell := \phi(h^j_\ell), \qquad f^j := \frac{1}{n}\langle v^j, h^j_L\rangle, \qquad g^j_\ell := n\nabla_{h^j_\ell} f^j.$$

For $j < k$, set $\chi^j := \partial_f \mathcal{L}(f^j, y^j)$. Write $X^{<j} = [x^0 \ \cdots \ x^{j-1}]$ and $q^{s,r}_x = d^{-1}\langle x^s, x^r\rangle$. For $\theta \in \{h, \phi, g\}$, define

$$\Theta^{<j}_\ell := [\theta^0_\ell \ \cdots \ \theta^{j-1}_\ell] \in \mathbb{R}^{n\times j}.$$

The loss coefficients are collected in

$$\chi^{<j} := (\chi^0, \dots, \chi^{j-1})^\top \in \mathbb{R}^j, \quad D^{<j}_\chi := \mathrm{diag}(\chi^{<j}) \in \mathbb{R}^{j\times j}.$$

The accumulated gradient updates are

$$U^j = U^0 - \frac{\eta_c}{\sqrt{d}} G^{<j}_0 D^{<j}_\chi (X^{<j})^\top, \tag{167}$$

$$W^j_\ell = W^0_\ell - \frac{\eta_c}{\sqrt{Ln}} G^{<j}_\ell D^{<j}_\chi (\Phi^{<j}_{\ell-1})^\top, \tag{168}$$

$$v^j = v^0 - \eta_c H^{<j}_L \chi^{<j}. \tag{169}$$

For ordered sets of training indices I, J, define

$$C^{I,J}_{\theta,\ell} := \left(\frac{1}{n}\langle \theta^s_\ell, \theta^r_\ell\rangle\right)_{s\in I, r\in J} \in \mathbb{R}^{|I|\times|J|}, \qquad \theta \in \{h, \phi, g\}.$$

Substituting Eq. (169) into the network gives the forward recursion

$$h^j_0 = \frac{1}{\sqrt{d}} U^0 x^j - \eta_c G^{<j}_0 D^{<j}_\chi q^{<j,j}_x, \tag{170}$$

$$h^j_\ell = h^j_{\ell-1} + \frac{1}{\sqrt{L}} a^j_\ell - \frac{\eta_c}{L} G^{<j}_\ell D^{<j}_\chi C^{<j,j}_{\phi,\ell-1}, \tag{171}$$

$$a^j_\ell := \frac{1}{\sqrt{n}} W^0_\ell \phi^j_{\ell-1}, \qquad 1 \le \ell \le L \tag{172}$$

and the backward recursion

$$g_L^j = v^j, \tag{173}$$

$$g_{\ell-1}^j = g_\ell^j + \phi'(h_{\ell-1}^j) \odot \left[\frac{1}{\sqrt{L}} b_\ell^j - \frac{\eta_c}{L} \Phi_{\ell-1}^{<j} D_\chi^{<j} C_{g,\ell}^{<j,j} \right], \tag{174}$$

$$b_\ell^j := \frac{1}{\sqrt{n}} (W_\ell^0)^\top g_\ell^j, \qquad \ell = L, \dots, 1. \tag{175}$$

Fixed-Depth Mean-Field Dynamics. We next describe the fixed-depth mean field for one representative coordinate, suppressing its coordinate index. The variables $\bar{h}_\ell^j, \bar{g}_\ell^j, \bar{a}_\ell^j, \bar{b}_\ell^j$ are scalars, and we use column vectors for their histories.

$$\bar{\theta}_\ell^I := (\bar{\theta}_\ell^s)_{s \in I} \in \mathbb{R}^{|I|}, \qquad \theta \in \{h, \phi, g\}, \qquad \bar{\phi}_\ell^j := \phi(\bar{h}_\ell^j).$$

The limiting second-moment kernels are

$$\bar{C}_{\theta,\ell}^{I,J} := \mathbb{E}\big[\bar{\theta}_\ell^I (\bar{\theta}_\ell^J)^\top\big] \in \mathbb{R}^{|I| \times |J|}.$$

Let $(\xi_0^j)_{j=0}^k$ be centered Gaussian with covariance $\mathbb{E}[\xi_0^s \xi_0^r] = q_x^{s,r}$, and let $v^0 \sim \mathcal{N}(0,1)$ denote the independent representative initial readout. The centered Gaussian fields at residual layer ℓ satisfy

$$\mathbb{E}[\xi_\ell^s \xi_\ell^r] = \bar{C}_{\phi,\ell-1}^{s,r}, \qquad \mathbb{E}[\zeta_\ell^s \zeta_\ell^r] = \bar{C}_{g,\ell}^{s,r}.$$

The forward and backward Gaussian families are mutually independent, independent across layers, and independent of the input Gaussian family and v^0. Their histories ξ_ℓ^I, ζ_ℓ^I are column vectors.

Define the mixed kernels

$$\bar{C}_{\zeta,\phi,\ell}^{I,J} := \mathbb{E}[\zeta_\ell^I (\bar{\phi}_{\ell-1}^J)^\top], \qquad \bar{C}_{\xi,g,\ell}^{I,J} := \mathbb{E}[\xi_\ell^I (\bar{g}_\ell^J)^\top].$$

Here ℓ identifies the weight block W_ℓ^0. The forward and backward response vectors are

$$\bar{R}_\ell^j := (\bar{R}_\ell^{j,0}, \dots, \bar{R}_\ell^{j,j-1})^\top = \sqrt{L} \left(\bar{C}_{g,\ell}^{<j,<j}\right)^\dagger \bar{C}_{\zeta,\phi,\ell}^{<j,j} \in \mathbb{R}^j, \tag{176}$$

$$\bar{S}_\ell^j := (\bar{S}_\ell^{j,0}, \dots, \bar{S}_\ell^{j,j})^\top = \sqrt{L} \left(\bar{C}_{\phi,\ell-1}^{\le j,\le j}\right)^\dagger \bar{C}_{\xi,g,\ell}^{\le j,j} \in \mathbb{R}^{j+1}, \tag{177}$$

where † denotes the Moore–Penrose inverse. The limiting reused-weight actions are

$$\bar{a}_\ell^j = \xi_\ell^j + \frac{1}{\sqrt{L}} (\bar{g}_\ell^{<j})^\top \bar{R}_\ell^j, \qquad \bar{b}_\ell^j = \zeta_\ell^j + \frac{1}{\sqrt{L}} (\bar{\phi}_{\ell-1}^{\le j})^\top \bar{S}_\ell^j. \tag{178}$$

The backward response includes the current forward feature because a_ℓ^j is evaluated before b_ℓ^j.

Set

$$\bar{f}^j := \mathbb{E}[\bar{v}^j \bar{h}_L^j], \qquad \bar{\chi}^j := \partial_f \mathcal{L}(\bar{f}^j, y^j) \quad (j < k), \qquad \bar{D}_\chi^{<j} := \mathrm{diag}(\bar{\chi}^{<j}),$$

where $\bar{\chi}^{<j}$ is a column vector. The mean-field recursions are

$$\bar{h}_0^j = \xi_0^j - \eta_c (\bar{g}_0^{<j})^\top \bar{D}_\chi^{<j} q_x^{<j,j}, \tag{179}$$

$$\bar{h}_\ell^j = \bar{h}_{\ell-1}^j + \frac{1}{\sqrt{L}} \bar{a}_\ell^j - \frac{\eta_c}{L} (\bar{g}_\ell^{<j})^\top \bar{D}_\chi^{<j} \bar{C}_{\phi,\ell-1}^{<j,j}, \tag{180}$$

$$\bar{g}_L^j = \bar{v}^j = v^0 - \eta_c (\bar{h}_L^{<j})^\top \bar{\chi}^{<j}, \tag{181}$$

$$\bar{g}_{\ell-1}^j = \bar{g}_\ell^j + \phi'(\bar{h}_{\ell-1}^j) \left[\frac{1}{\sqrt{L}} \bar{b}_\ell^j - \frac{\eta_c}{L} (\bar{\phi}_{\ell-1}^{<j})^\top \bar{D}_\chi^{<j} \bar{C}_{g,\ell}^{<j,j} \right]. \tag{182}$$

At each iterate, the construction proceeds through the forward pass, readout, and backward pass. Different mean-field coordinates are independent copies of this joint process.

Continuous-depth mean-field dynamics. We couple the discrete Gaussian fields and the continuous-depth noises through the same Brownian motions. At each depth, the coupling uses Gaussian regression in training order.

Let $Z^{\le j}$ be a centered Gaussian history with covariance blocks $C^{I,J}$. If $A_{<j,<j}$ is the lower triangular factor constructed for $C^{<j,<j}$, the conditional mean and variance of Z^j given $Z^{<j}$ give

$$A_{j,<j} = C^{j,<j}(C^{<j,<j})^\dagger A_{<j,<j}, \qquad A_{jj} = \left(C^{j,j} - C^{j,<j}(C^{<j,<j})^\dagger C^{<j,j}\right)^{1/2}. \tag{183}$$

At $j = 0$, the older block is empty and $A_{00} = \sqrt{C^{0,0}}$. The construction also applies to singular covariance blocks. Positive semidefiniteness gives $C^{<j,j} \in \mathrm{ran}(C^{<j,<j})$, so the residual is independent of the older Gaussian history. Its variance may vanish.

Apply Eq. (183) to the discrete kernels

$$\sigma_\ell^L(\sigma_\ell^L)^\top = \bar{C}_{\phi,\ell-1}^{\le k,\le k}, \qquad \tau_\ell^L(\tau_\ell^L)^\top = \bar{C}_{g,\ell}^{\le k,\le k} \tag{184}$$

to obtain the factors σ_ℓ^L and τ_ℓ^L, respectively. Take mutually independent standard Brownian motions $B^0, \dots, B^k, \widetilde{B}^0, \dots, \widetilde{B}^k$, independent of the input Gaussian family and v^0. Write $B_t^{<j} = (B_t^0, \dots, B_t^{j-1})^\top$ and define $\widetilde{B}_t^{<j}$ analogously. On the grid $t_\ell = \ell\Delta$ with $\Delta = 1/L$, write $\Delta B_\ell^j = B_{t_\ell}^j - B_{t_{\ell-1}}^j$ and use the same increment notation for the history vectors and backward drivers. The discrete Gaussian innovations are realized as

$$\sqrt{\Delta}\,\xi_\ell^{j,L} = \sigma_{\ell,j,<j}^L\,\Delta B_\ell^{<j} + \sigma_{\ell,jj}^L\,\Delta B_\ell^j, \tag{185}$$

$$\sqrt{\Delta}\,\zeta_\ell^{j,L} = \tau_{\ell,j,<j}^L\,\Delta \widetilde{B}_\ell^{<j} + \tau_{\ell,jj}^L\,\Delta \widetilde{B}_\ell^j. \tag{186}$$

In each line, the first term gives the conditional mean from the older increments and the second gives the independent residual.

For the continuous-depth equations, write H_t^j, G_t^j for the forward and backward states. Their histories H_t^I, G_t^I are column vectors. Define

$$\phi_t^j := \phi(H_t^j), \qquad C_\phi^{I,J}(t) := \mathbb{E}[\phi_t^I(\phi_t^J)^\top], \qquad C_g^{I,J}(t) := \mathbb{E}[G_t^I(G_t^J)^\top].$$

Applying the same row construction pointwise in depth gives deterministic factors $\sigma(t)$ and $\tau(t)$ of $C_\phi^{\le k,\le k}(t)$ and $C_g^{\le k,\le k}(t)$, respectively. For measurable kernels with

$$\int_0^1 \left[\mathrm{tr}\, C_\phi^{\le k,\le k}(t) + \mathrm{tr}\, C_g^{\le k,\le k}(t)\right] \mathrm{d}t < \infty,$$

the corresponding continuous noises are the Gaussian martingales

$$W_t^j := \int_0^t \sigma_{j,<j}(u)\,\mathrm{d}B_u^{<j} + \int_0^t \sigma_{jj}(u)\,\mathrm{d}B_u^j, \tag{187}$$

$$\widetilde{W}_t^j := \int_0^t \tau_{j,<j}(u)\,\mathrm{d}\widetilde{B}_u^{<j} + \int_0^t \tau_{jj}(u)\,\mathrm{d}\widetilde{B}_u^j. \tag{188}$$

Their quadratic covariations recover the kernels,

$$\mathrm{d}[W^j, W^r]_t = C_\phi^{j,r}(t)\,\mathrm{d}t, \qquad \mathrm{d}[\widetilde{W}^j, \widetilde{W}^r]_t = C_g^{j,r}(t)\,\mathrm{d}t, \qquad \mathrm{d}[W^j, \widetilde{W}^r]_t = 0.$$

In particular,

$$\mathbb{E}[W_t^j W_u^r] = \int_0^{t\wedge u} C_\phi^{j,r}(s)\,\mathrm{d}s, \qquad \mathbb{E}[\widetilde{W}_t^j \widetilde{W}_u^r] = \int_0^{t\wedge u} C_g^{j,r}(s)\,\mathrm{d}s.$$

The two noise families are independent. Within each family, the components are generally correlated across training steps.

Let $R^j(t) \in \mathbb{R}^j$ and $S^j(t) \in \mathbb{R}^{j+1}$ denote the forward response and the candidate backward response. Define the readout and loss coefficients in training order by

$$V^j := v^0 - \eta_c (H_1^{<j})^\top \chi_\infty^{<j}, \qquad f_\infty^j := \mathbb{E}[V^j H_1^j], \tag{189}$$

$$\chi_\infty^j := \partial_f \mathcal{L}(f_\infty^j, y^j) \quad (j < k), \qquad D_{\chi_\infty}^{<j} := \mathrm{diag}(\chi_\infty^{<j}). \tag{190}$$

The drift terms are

$$F^j(t) := (G_t^{<j})^\top \big[R^j(t) - \eta_c D_{\chi_\infty}^{<j} C_\phi^{<j,j}(t)\big], \tag{191}$$

$$K^j(t) := (\phi_t^{\leq j})^\top S^j(t) - \eta_c (\phi_t^{<j})^\top D_{\chi_\infty}^{<j} C_g^{<j,j}(t). \tag{192}$$

The continuous-depth equations take the compact form

$$\mathrm{d}H_t^j = F^j(t)\,\mathrm{d}t + \mathrm{d}W_t^j, \tag{193}$$

$$\mathrm{d}G_t^j = -\phi'(H_t^j)\big[K^j(t)\,\mathrm{d}t + \mathrm{d}\widetilde{W}_t^j\big], \tag{194}$$

with boundary conditions

$$H_0^j = \xi_0^j - \eta_c (G_0^{<j})^\top D_{\chi_\infty}^{<j} q_x^{<j,j}, \qquad G_1^j = V^j.$$

Substituting Eq. (188) gives the equivalent display with the older and fresh drivers separated,

$$\mathrm{d}H_t^j = F^j(t)\,\mathrm{d}t + \sigma_{j,<j}(t)\,\mathrm{d}B_t^{<j} + \sigma_{jj}(t)\,\mathrm{d}B_t^j, \tag{195}$$

$$\mathrm{d}G_t^j = -\phi'(H_t^j)\left[K^j(t)\,\mathrm{d}t + \tau_{j,<j}(t)\,\mathrm{d}\widetilde{B}_t^{<j} + \tau_{jj}(t)\,\mathrm{d}\widetilde{B}_t^j\right]. \tag{196}$$

The forward noise has deterministic coefficients and is integrated in the Itô sense. The backward equation is formal and records the discrete convention in which $\phi'(\bar{h}_{\ell-1}^{j,L})$ multiplies the layer-ℓ increment. After training, the forward feature can depend on the shared older backward noise, so the stochastic integral in the physical backward equation requires a separate justification. The minus sign expresses the backward propagation from the terminal condition at depth 1. For linear activation, Subsection F.3 gives the response closure and the limiting equations with deterministic Brownian integrands.

F.2 Width Convergence at Fixed Depth

F.2.1 Assumptions and proof strategy

Fix $L \geq 1$ and a finite number k of updates. We use the finite and limiting dynamics defined above, suppressing n, L when convenient.

Throughout this subsection, we impose the following assumptions.

Assumption F.1. Fix $L \geq 1$, $k < \infty$, and $\eta_c \geq 0$.

1. The inputs $x^0, \ldots, x^k \in \mathbb{R}^d$ and training labels $y^0, \ldots, y^{k-1}$ are deterministic, with $\|x^0\|^2/d > 0$. The entries of U^0, W_ℓ^0, and v^0 are independent standard Gaussians.
2. The activation $\phi \in C^1(\mathbb{R})$ is globally Lipschitz, and ϕ' is globally Lipschitz. In particular, $\|\phi'\|_\infty < \infty$.
3. For each $s < k$, the map $u \mapsto \partial_f \mathcal{L}(u, y^s)$ is globally Lipschitz and finite at zero.
4. The recursively constructed mean field satisfies, for some $\kappa > 0$,

$$\min_{1 \leq \ell \leq L} \left\{ \lambda_{\min}\big(\bar{C}_{\phi,\ell-1}^{\leq k,\leq k}\big), \lambda_{\min}\big(\bar{C}_{g,\ell}^{\leq k,\leq k}\big) \right\} \geq \kappa. \tag{197}$$

The constant κ may depend on the fixed L, k. The nondegeneracy condition applies to the feature and gradient Gram matrices. The constants $C_{p,L,k}$ are independent of n and may depend on the data, η_c, the activation and loss constants, and κ^{-1}. Matrix L^p norms use the operator norm.

The proof extends Proposition E.4. We establish limiting moments, control empirical coefficients on good events, and use the query estimates to close the induction over training iterates and layers. Condition Eq. (197) replaces the scalar nondegeneracy estimates used there.

F.2.2 Moment bounds for the limiting process

Lemma F.2 (Moments of the limiting multistep process)**.** *Suppose Assumption F.1 holds. Then, for every $p \geq 2$,*

$$\max_{0 \leq j \leq k} \left[\max_{0 \leq \ell \leq L} \left(\|\bar{h}_\ell^j\|_{L^p} + \|\bar{g}_\ell^j\|_{L^p} \right) + \max_{1 \leq \ell \leq L} \left(\|\bar{a}_\ell^j\|_{L^p} + \|\bar{b}_\ell^j\|_{L^p} \right) \right] \leq C_{p,L,k}. \tag{198}$$

The joint limiting coordinate processes are iid across coordinates.

Proof. Initialization and training induction. The initialization moment bounds in Lemma D.1 give the base case $j = 0$. At this iterate, the weight actions are the Gaussian innovations by Eq. (178), so their finite moments follow from the kernel bounds. The initial output is finite by Cauchy–Schwarz, so $\bar{\chi}^0$ is finite when $k \geq 1$. Suppose that all limiting states and weight actions through step $j-1$ have moments of every finite order, and that $\bar{\chi}^0, \dots, \bar{\chi}^{j-1}$ are finite.

By Eq. (180) and Eq. (182), the input state $\bar{h}_0^j$ and readout $\bar{v}^j$ involve only Gaussian inputs, the initial readout, and previous trajectories with deterministic coefficients. Hence both have moments of every finite order.

Forward states and actions. Induct forward over layers. Suppose that $\bar{h}_{\ell-1}^j$ has all finite moments. The Lipschitz growth of ϕ gives the same property for $\bar{\phi}_{\ell-1}^j$, and Cauchy–Schwarz gives

$$|\bar{C}_{\phi,\ell-1}^{s,j}| \leq \|\bar{\phi}_{\ell-1}^s\|_{L^2} \|\bar{\phi}_{\ell-1}^j\|_{L^2}, \qquad s < j.$$

Every history Gram matrix appearing in a response is a principal submatrix of a matrix in Eq. (197), so its inverse has operator norm at most κ^{-1}. The forward response definition Eq. (177) and Cauchy–Schwarz give

$$\|\bar{R}_\ell^j\| \leq \frac{\sqrt{L}}{\kappa} \left(\operatorname{tr} \bar{C}_{g,\ell}^{<j,<j}\right)^{1/2} \|\bar{\phi}_{\ell-1}^j\|_{L^2} < \infty. \tag{199}$$

Here $\mathbb{E}[\|\zeta_\ell^{<j}\|^2] = \operatorname{tr} \bar{C}_{g,\ell}^{<j,<j}$, and the gradient history is controlled by the training-step induction hypothesis. The Gaussian variable ξ_ℓ^j has variance $\mathbb{E}[(\bar{\phi}_{\ell-1}^j)^2]$, so it has all finite moments. Thus Eq. (178) gives all finite moments of $\bar{a}_\ell^j$. The forward recursion then gives the same property for $\bar{h}_\ell^j$, completing the forward layer induction.

The network output is finite because

$$|\bar{f}^j| = |\mathbb{E}[\bar{v}^j \bar{h}_L^j]| \leq \|\bar{v}^j\|_{L^2} \|\bar{h}_L^j\|_{L^2} < \infty.$$

When $j < k$, the loss assumption consequently gives $|\bar{\chi}^j| < \infty$.

Backward states and actions. We next induct backward over layers, starting from $\bar{g}_L^j = \bar{v}^j$. Suppose that $\bar{g}_\ell^j$ has all finite moments. All forward features through step j have already been constructed. Therefore,

$$|\bar{C}_{g,\ell}^{s,j}| \leq \|\bar{g}_\ell^s\|_{L^2} \|\bar{g}_\ell^j\|_{L^2}, \qquad s < j,$$

and the backward response satisfies

$$\|\bar{S}_\ell^j\| \leq \frac{\sqrt{L}}{\kappa} \left(\operatorname{tr} \bar{C}_{\phi,\ell-1}^{\leq j,\leq j}\right)^{1/2} \|\bar{g}_\ell^j\|_{L^2} < \infty. \tag{200}$$

Since ζ_ℓ^j is Gaussian with variance $\mathbb{E}[(\bar{g}_\ell^j)^2]$, it has all finite moments. Eq. (178) therefore gives all finite moments of $\bar{b}_\ell^j$. Finally, boundedness of ϕ' and the backward recursion give all finite moments of $\bar{g}_{\ell-1}^j$. This closes the backward layer induction and the induction over training steps. Taking the maximum over the finitely many steps and layers gives Eq. (198). The expectation coefficients are deterministic, and each limiting coordinate process is the same function of an independent copy of the primitive Gaussian family. Hence the coordinate processes are iid. □

F.2.3 Approximation of the reused-weight actions

Lemma F.3 (Approximation of the reused-weight actions)**.** *Suppose Assumption F.1 holds. Fix $1 \leq j \leq k$ and $1 \leq \ell \leq L$. Use the successive Gaussian coupling extending the initialization coupling.*

For the forward query, suppose that, for every finite $q \geq 2$,

$$\sup_i \|u_i - \bar{u}_i\|_{L^q} \leq C_{q,L,k} n^{-1/2}, \qquad \sup_i \|u_i\|_{L^q} \leq C_{q,L,k}, \tag{201}$$

for

$$u \in \{h_{\ell-1}^j\} \cup \{\phi_{\ell-1}^s, g_\ell^s, a_\ell^s, b_\ell^s : 0 \leq s < j\}.$$

Then, for every $p \geq 2$,

$$\|C_{\phi,\ell-1}^{\leq j,\leq j} - \bar{C}_{\phi,\ell-1}^{\leq j,\leq j}\|_{L^p} \leq C_{p,L,k} n^{-1/2}, \tag{202}$$

$$\sup_i \|a_{\ell,i}^j - \bar{a}_{\ell,i}^j\|_{L^p} \leq C_{p,L,k} n^{-1/2}. \tag{203}$$

For the backward query, suppose instead that Eq. (201) holds for

$$u \in \{\phi_{\ell-1}^s, a_\ell^s : 0 \leq s \leq j\} \cup \{g_\ell^s, b_\ell^s : 0 \leq s < j\} \cup \{g_\ell^j\}.$$

Then

$$\|C_{g,\ell}^{\leq j,\leq j} - \bar{C}_{g,\ell}^{\leq j,\leq j}\|_{L^p} \leq C_{p,L,k} n^{-1/2}, \tag{204}$$

$$\sup_i \|b_{\ell,i}^j - \bar{b}_{\ell,i}^j\|_{L^p} \leq C_{p,L,k} n^{-1/2}. \tag{205}$$

In either case, the new finite-width query has coordinate moments of every finite order bounded uniformly in n.

Proof. Coupling and empirical products. Use the joint Gaussian realization from Lemma C.1, sharing each fresh innovation between the finite and limiting queries.

The induction hypotheses, Lemma F.2, and Eq. (25) give

$$\left\| \frac{1}{n} \sum_i u_i v_i - \mathbb{E}[\bar{u}_i \bar{v}_i] \right\|_{L^q} \leq C_{q,L,k} n^{-1/2} \tag{206}$$

for any available pair of coordinate variables and every finite $q \geq 2$. For the current forward feature, the required estimates follow from the Lipschitz property of ϕ and the hypothesis on $h_{\ell-1}^j$. Since the history dimension is bounded independently of n, the entrywise estimates for $C_{\phi,\ell-1}^{\leq j,\leq j}$ prove Eq. (202).

Good events and Gram inverses. We first consider the forward query. Define

$$\Omega_a := \left\{ \lambda_{\min}(C_{\phi,\ell-1}^{<j,<j}) \geq \kappa/2, \quad \lambda_{\min}(C_{g,\ell}^{<j,<j}) \geq \kappa/2 \right\}.$$

By Assumption F.1, both limiting history Gram matrices have smallest eigenvalue at least κ. Thus, on Ω_a^c, at least one empirical Gram matrix differs from its limit by at least $\kappa/2$ in operator norm. Applying Eq. (206) at order $2r$ and using Markov's inequality gives

$$\mathbb{P}(\Omega_a^c) \leq C_{r,L,k} n^{-r}, \qquad r \geq 1. \tag{207}$$

On Ω_a, the two empirical Gram inverses have operator norm at most $2/\kappa$. For either Gram matrix, the identity $C^{-1} - \bar{C}^{-1} = C^{-1}(\bar{C} - C)\bar{C}^{-1}$ and Eq. (206) therefore give

$$\|1_{\Omega_a}(C^{-1} - \bar{C}^{-1})\|_{L^q} \leq C_{q,L,k} n^{-1/2}. \tag{208}$$

Identification of the limiting query. On Ω_a, the pseudoinverses in Eq. (36) are ordinary inverses. The limiting counterpart of the finite residual in Eq. (32) is

$$\bar{\phi}_{\ell-1,i,\perp}^j := \bar{\phi}_{\ell-1,i}^j - (\bar{\phi}_{\ell-1,i}^{<j})^\top (\bar{C}_{\phi,\ell-1}^{<j,<j})^{-1} \bar{C}_{\phi,\ell-1}^{<j,j}.$$

To identify the limiting regression terms, use Gaussian regression for ξ_ℓ^j and the overlap identity

$$\mathbb{E}[\zeta_\ell^r \bar{\phi}_{\ell-1}^s] = \frac{1}{\sqrt{L}} \sum_{t<s} \bar{C}_{g,\ell}^{r,t} \bar{R}_\ell^{s,t}, \qquad r, s < j.$$

For $r < s$, this follows from the response definition Eq. (177). For $r \geq s$, regress ζ_ℓ^r onto its history before step s. The remaining Gaussian component is independent of $\bar{\phi}_{\ell-1}^s$, which is constructed before the backward innovations at step s. Moreover, $\bar{\phi}_{\ell-1,i,\perp}^j$ is orthogonal in L^2 to the feature history. Thus, in the mixed product with this residual, the response part of $\bar{b}_{\ell,i}^{<j}$ in Eq. (178) vanishes.

The overlap identity then cancels the inherited responses from $\bar{a}_{\ell,i}^{<j}$, leaving the current response $(\bar{g}_{\ell,i}^{<j})^\top \bar{R}_\ell^j/\sqrt{L}$.

Decomposition and estimation of the query error. Subtracting the resulting limiting regression from Eq. (36) gives, on Ω_a,

$$\begin{aligned} a_{\ell,i}^j - \bar{a}_{\ell,i}^j = &\left[(A_\ell^{<j})_{i,:}(C_{\phi,\ell-1}^{<j,<j})^{-1}C_{\phi,\ell-1}^{<j,j} - (\bar{a}_{\ell,i}^{<j})^\top(\bar{C}_{\phi,\ell-1}^{<j,<j})^{-1}\bar{C}_{\phi,\ell-1}^{<j,j}\right] \\ &+ \left[(G_\ell^{<j})_{i,:}(C_{g,\ell}^{<j,<j})^{-1}\frac{1}{n}(B_\ell^{<j})^\top\phi_{\ell-1,\perp}^j - (\bar{g}_{\ell,i}^{<j})^\top(\bar{C}_{g,\ell}^{<j,<j})^{-1}\mathbb{E}[\bar{b}_{\ell,i}^{<j}\bar{\phi}_{\ell-1,i,\perp}^j]\right] \\ &+ \left[\left(\frac{\|\phi_{\ell-1,\perp}^j\|}{\sqrt{n}} - \sqrt{\mathbb{E}[(\bar{\phi}_{\ell-1,i,\perp}^j)^2]}\right)z_{\ell,i}^j - \frac{\|\phi_{\ell-1,\perp}^j\|}{\sqrt{n}}(P_{G_\ell^{<j}}z_\ell^j)_i\right]. \end{aligned} \tag{209}$$

The three brackets are the forward-history, response, and Gaussian errors, respectively.

For the first bracket, the previous a-query estimates, Eq. (206), and Eq. (208) give an $O(n^{-1/2})$ error in L^p on Ω_a, by the same product decomposition and Hölder estimate as in Lemma E.3.

For the second and third brackets, first compare the residual features. Applying the incoming feature estimates, Eq. (206), and Eq. (208) to their definitions gives

$$\sup_i \|1_{\Omega_a}(\phi_{\ell-1,i,\perp}^j - \bar{\phi}_{\ell-1,i,\perp}^j)\|_{L^q} \le C_{q,L,k}n^{-1/2}. \tag{210}$$

Together with the limiting residual moment bounds, this also gives uniformly bounded coordinate L^q moments of the finite residual on Ω_a.

For the second bracket, Eq. (210), the previous b-query estimates, and the empirical-product argument in Eq. (25) give

$$\left\|1_{\Omega_a}\left[\frac{1}{n}(B_\ell^{<j})^\top\phi_{\ell-1,\perp}^j - \mathbb{E}[\bar{b}_{\ell,i}^{<j}\bar{\phi}_{\ell-1,i,\perp}^j]\right]\right\|_{L^q} \le C_{q,L,k}n^{-1/2}. \tag{211}$$

The coupled product errors are estimated on Ω_a, and the centered iid limiting average is estimated before multiplication by 1_{Ω_a}. Combining Eq. (211) with Eq. (208) and the previous gradient-state estimates bounds the second bracket by $C_{p,L,k}n^{-1/2}$ in L^p on Ω_a.

For the third bracket, apply the same empirical-product argument to the residual squares, using Eq. (210) and the iid limiting moment bounds from Lemma F.2. This gives

$$\left\|1_{\Omega_a}\left[\frac{\|\phi_{\ell-1,\perp}^j\|^2}{n} - \mathbb{E}[(\bar{\phi}_{\ell-1,i,\perp}^j)^2]\right]\right\|_{L^q} \le C_{q,L,k}n^{-1/2}. \tag{212}$$

The limiting residual variance is the Schur complement in $\bar{C}_{\phi,\ell-1}^{\le j,\le j} \succeq \kappa I$ and is therefore at least κ. Thus $|\sqrt{x} - \sqrt{y}| \le |x-y|/\sqrt{y}$ and Eq. (212) give

$$\left\|1_{\Omega_a}\left[\frac{\|\phi_{\ell-1,\perp}^j\|}{\sqrt{n}} - \sqrt{\mathbb{E}[(\bar{\phi}_{\ell-1,i,\perp}^j)^2]}\right]\right\|_{L^q} \le C_{q,L,k}n^{-1/2}. \tag{213}$$

Multiplication by the independent $z_{\ell,i}^j$ preserves this rate.

For the projection correction, on Ω_a,

$$(P_{G_\ell^{<j}})_{ii} \le \frac{2}{\kappa n}\sum_{s<j}|g_{\ell,i}^s|^2.$$

As in Lemma E.3, the conditional Gaussian variance of $(P_{G_\ell^{<j}}z_\ell^j)_i$ is $(P_{G_\ell^{<j}})_{ii}$. Using this bound, the incoming coordinate moments, and $\|\phi_{\ell-1,\perp}^j\| \le \|\phi_{\ell-1}^j\|$ gives

$$\sup_i \left\|1_{\Omega_a}\frac{\|\phi_{\ell-1,\perp}^j\|}{\sqrt{n}}(P_{G_\ell^{<j}}z_\ell^j)_i\right\|_{L^p} \le C_{p,L,k}n^{-1/2}.$$

Combining the three brackets proves

$$\sup_i \|1_{\Omega_a}(a^j_{\ell,i} - \bar{a}^j_{\ell,i})\|_{L^p} \le C_{p,L,k} n^{-1/2}. \tag{214}$$

Bad event and the backward query. On the bad event, use the original forward query definition in Eq. (172). The Gaussian operator-norm bound and the incoming feature moments give $\sup_i \|a^j_{\ell,i}\|_{L^{2p}} \le C_{p,L,k}\sqrt{n}$, as in Lemma E.3. Using Eq. (207) with $r = 2p$ and the limiting moments from Lemma F.2, we obtain

$$\sup_i \|1_{\Omega_a^c}(a^j_{\ell,i} - \bar{a}^j_{\ell,i})\|_{L^p} \le \mathbb{P}(\Omega_a^c)^{1/(2p)} \sup_i \|a^j_{\ell,i} - \bar{a}^j_{\ell,i}\|_{L^{2p}} \le C_{p,L,k} n^{-1/2}.$$

Together with Eq. (214), this proves Eq. (203). Adding the error to the limiting moment bound gives $\sup_i \|a^j_{\ell,i}\|_{L^p} \le C_{p,L,k}$.

The same argument applied to Eq. (37) under the backward induction hypotheses proves Eq. (204) and Eq. (205), together with $\sup_i \|b^j_{\ell,i}\|_{L^p} \le C_{p,L,k}$. □

F.2.4 Width convergence over finitely many updates

Proposition F.4 (Width convergence over finitely many SGD updates)**.** *Suppose Assumption F.1 holds. In the joint Gaussian coupling, for every $p \ge 2$,*

$$\max_{\substack{0\le j\le k\\ 0\le \ell\le L}} \sup_i \left(\|h^j_{\ell,i} - \bar{h}^j_{\ell,i}\|_{L^p} + \|g^j_{\ell,i} - \bar{g}^j_{\ell,i}\|_{L^p}\right) \le C_{p,L,k} n^{-1/2}, \tag{215}$$

$$\max_{\substack{0\le j\le k\\ 1\le \ell\le L}} \sup_i \left(\|a^j_{\ell,i} - \bar{a}^j_{\ell,i}\|_{L^p} + \|b^j_{\ell,i} - \bar{b}^j_{\ell,i}\|_{L^p}\right) \le C_{p,L,k} n^{-1/2}. \tag{216}$$

The finite-width states and queries have uniformly bounded coordinate L^p moments. Moreover,

$$\max_{0\le j\le k} \|f^j - \bar{f}^j\|_{L^p} \le C_{p,L,k} n^{-1/2}, \tag{217}$$

$$\|\chi^j - \bar{\chi}^j\|_{L^p} \le C_{p,L,k} n^{-1/2}, \qquad 0 \le j < k. \tag{218}$$

For $\theta \in \{h, \phi, g\}$, the history kernels satisfy

$$\max_{0\le \ell\le L} \|C^{\le k,\le k}_{\theta,\ell} - \bar{C}^{\le k,\le k}_{\theta,\ell}\|_{L^p} \le C_{p,L,k} n^{-1/2}. \tag{219}$$

Proof. Initialization. At $j = 0$, the state estimates follow from Propositions D.2 and D.5, and the query estimates follow from Corollary D.6. Corollary D.7 gives the output estimate. When $k \ge 1$, Lipschitz continuity of the loss derivative gives the corresponding estimate for χ^0. The coordinate estimates and the limiting moment bounds give uniform moments at initialization.

Input and forward pass. We induct over training iterates, carrying the estimates at all finite moment orders simultaneously. Suppose the estimates hold through iterate $j - 1$.

The input recursions in Eq. (172) and Eq. (180) share the same Gaussian input. Their remaining terms involve only previous loss derivatives and backward states. The training induction hypothesis and Hölder's inequality therefore give the $O(n^{-1/2})$ estimate for $h^j_0 - \bar{h}^j_0$. Together with the limiting moment bounds, this gives uniformly bounded coordinate moments for h^j_0.

We then proceed forward over the layers. Suppose the estimates hold for $h^j_{\ell-1}$. Together with the estimates from earlier training iterates, this gives the forward hypotheses of Lemma F.3. Its conclusions control $a^j_\ell - \bar{a}^j_\ell$ and the feature kernel in the SGD correction. Comparing Eq. (172) and Eq. (180), the query error and the product errors in the SGD correction give the estimate for h^j_ℓ. Uniform coordinate moments follow by adding the error to the limiting moments from Lemma F.2. This closes the forward layer induction.

Readout, output, and loss. The readout updates use only the previous terminal states $h^s_L, \bar{h}^s_L$ and loss derivatives $\chi^s, \bar{\chi}^s$, with $s < j$. Thus the training induction hypothesis gives the $O(n^{-1/2})$ coordinate

estimate for $v^j - \bar{v}^j$. The limiting moments of $\bar{v}^j = \bar{g}_L^j$ then give uniformly bounded coordinate moments for v^j. With these readout and forward estimates, Eq. (25) applied to

$$f^j = \frac{1}{n}\sum_i v_i^j h_{L,i}^j, \qquad \bar{f}^j = \mathbb{E}[\bar{v}^j \bar{h}_L^j]$$

gives the estimate in Eq. (217) at step j. For $j < k$, the Lipschitz loss derivative then gives Eq. (218).

Backward pass and kernel estimates. Starting from $g_L^j = v^j$, the same layer induction in reverse order, using the backward part of Lemma F.3 and the boundedness and Lipschitz continuity of ϕ', gives the backward estimates. This closes the induction over training iterates.

Finally, apply Eq. (25) to the completed h and g histories and to the activated features $\phi(h)$, whose coordinate estimates follow from the Lipschitz property of ϕ. Since the history dimension is fixed, the entrywise estimates give Eq. (219). □

Remark F.5. The nondegeneracy assumption controls the Gram inverses and innovation square roots over the finitely many queries at fixed L, k. A retained independent history offers an approach to dependent limiting queries. Extending the proof in this way requires controlling the finite-width residual directions outside the retained history.

F.3 Depth Convergence for Linear Activation

Fix k SGD updates and suppose that items 1 and 3 of Assumption F.1 hold. Set $\phi(h) = h$, $\Delta = L^{-1}$, and $t_\ell = \ell\Delta$. Then $\bar{C}_{\phi,\ell} = \bar{C}_{h,\ell}$ and $\phi' \equiv 1$.

F.3.1 Fixed-depth mean field and response closure

Substitute the reused-weight actions $\bar{a}_\ell^j, \bar{b}_\ell^j$ from Eq. (178) into the forward and backward equations Eq. (180) and Eq. (182). For $\phi(h) = h$, the resulting scalar equations and boundary conditions are

$$\begin{aligned}
\bar{h}_0^j &= \xi_0^j - \eta_c(\bar{g}_0^{<j})^\top \bar{D}_\chi^{<j} q_x^{<j,j}, \\
\bar{h}_\ell^j &= \bar{h}_{\ell-1}^j + \sqrt{\Delta}\,\xi_\ell^j + \Delta(\bar{g}_\ell^{<j})^\top \big(\bar{R}_\ell^j - \eta_c \bar{D}_\chi^{<j} \bar{C}_{h,\ell-1}^{<j,j}\big), \\
\bar{g}_L^j = \bar{v}^j &= v^0 - \eta_c(\bar{h}_L^{<j})^\top \bar{\chi}^{<j}, \\
\bar{g}_{\ell-1}^j &= \bar{g}_\ell^j + \sqrt{\Delta}\,\zeta_\ell^j + \Delta(\bar{h}_{\ell-1}^{\le j})^\top \bar{S}_\ell^j - \eta_c \Delta(\bar{h}_{\ell-1}^{<j})^\top \bar{D}_\chi^{<j} \bar{C}_{g,\ell}^{<j,j}.
\end{aligned} \tag{220}$$

Here $\bar{R}_\ell^j \in \mathbb{R}^j$ and $\bar{S}_\ell^j \in \mathbb{R}^{j+1}$ are the mixed-moment responses in Eq. (177). In particular, the last equation initially includes $\bar{h}_{\ell-1}^j$ in the response product $(\bar{h}_{\ell-1}^{\le j})^\top \bar{S}_\ell^j$.

We seek strictly lower triangular matrices $\mathbf{R}_\ell, \mathbf{S}_\ell \in \mathbb{R}^{(k+1)\times(k+1)}$ whose rows give the same response products. Define the drift matrices $\mathbf{F}_\ell, \mathbf{K}_\ell \in \mathbb{R}^{(k+1)\times(k+1)}$ entrywise for $j, a \in \{0, \dots, k\}$ by

$$(\mathbf{F}_\ell)_{ja} = \begin{cases} (\mathbf{R}_\ell)_{ja} - \eta_c \bar{\chi}^a \bar{C}_{h,\ell-1}^{a,j}, & a < j, \\ 0, & a \ge j, \end{cases} \quad (\mathbf{K}_\ell)_{ja} = \begin{cases} (\mathbf{S}_\ell)_{ja} - \eta_c \bar{\chi}^a \bar{C}_{g,\ell}^{a,j}, & a < j, \\ 0, & a \ge j. \end{cases} \tag{221}$$

At depth ℓ, stack the training iterates into the columns

$$\bar{\mathbf{h}}_\ell = (\bar{h}_\ell^0, \dots, \bar{h}_\ell^k)^\top \in \mathbb{R}^{k+1}, \qquad \bar{\mathbf{g}}_\ell = (\bar{g}_\ell^0, \dots, \bar{g}_\ell^k)^\top \in \mathbb{R}^{k+1}.$$

Likewise, write $\boldsymbol{\xi}_\ell = (\xi_\ell^0, \dots, \xi_\ell^k)^\top$ and $\boldsymbol{\zeta}_\ell = (\zeta_\ell^0, \dots, \zeta_\ell^k)^\top$. The proposed vector recursions are

$$\bar{\mathbf{h}}_\ell = \bar{\mathbf{h}}_{\ell-1} + \sqrt{\Delta}\,\boldsymbol{\xi}_\ell + \Delta \mathbf{F}_\ell \bar{\mathbf{g}}_\ell, \qquad \bar{\mathbf{g}}_{\ell-1} = \bar{\mathbf{g}}_\ell + \sqrt{\Delta}\,\boldsymbol{\zeta}_\ell + \Delta \mathbf{K}_\ell \bar{\mathbf{h}}_{\ell-1}. \tag{222}$$

Thus $\bar{h}_\ell^j, \bar{g}_\ell^j$ are scalars, while $\bar{\mathbf{h}}_\ell, \bar{\mathbf{g}}_\ell \in \mathbb{R}^{k+1}$ stack training iterates at a single depth ℓ.

Hold the coefficients and Gaussian variables fixed. To define $\mathbf{H}_m$, give $\bar{\mathbf{g}}_m$ a value at depth m. Run Eq. (222) on layers $1, \dots, m$ in training order $j = 0, \dots, k$. At iterate j, use the input equation for $\bar{h}_0^j$ in Eq. (220) to start the forward pass, and the supplied $\bar{g}_m^j$ to start the backward pass. The backward pass supplies the gradients used by later iterates. To define $\mathbf{J}_m$, give $\bar{\mathbf{h}}_m$ a value at depth m and run the updates on layers $m+1, \dots, L$ in the same training order. At each iterate, start the forward pass from the supplied $\bar{h}_m^j$ and the backward pass from the readout equation for $\bar{g}_L^j$ in Eq. (220).

Define $\mathbf{H}_m, \mathbf{J}_m \in \mathbb{R}^{(k+1)\times(k+1)}$ by

$$(\mathbf{H}_m)_{ja} = \begin{cases} \dfrac{\partial \bar{h}_m^j}{\partial \bar{g}_m^a}, & a < j, \\ 0, & a \geq j, \end{cases} \qquad (\mathbf{J}_m)_{ja} = \begin{cases} \dfrac{\partial \bar{g}_m^j}{\partial \bar{h}_m^a}, & a < j, \\ 0, & a \geq j. \end{cases} \tag{223}$$

At $m = 0$ and $m = L$, the input and readout equations give, respectively,

$$\begin{aligned} \bar{\mathbf{h}}_0 &= \boldsymbol{\xi}_0 + \mathbf{H}_0 \bar{\mathbf{g}}_0, & (\mathbf{H}_0)_{ja} &= -\eta_c \bar{\chi}^a q_x^{a,j}, \\ \bar{\mathbf{g}}_L &= v^0 \mathbf{1} + \mathbf{J}_L \bar{\mathbf{h}}_L, & (\mathbf{J}_L)_{ja} &= -\eta_c \bar{\chi}^a, \qquad a < j. \end{aligned} \tag{224}$$

Lemma F.6 (Linear response recursions)**.** *The Jacobians $\mathbf{H}_\ell, \mathbf{J}_\ell$ satisfy*

$$\mathbf{H}_\ell = \mathbf{H}_{\ell-1}(\mathbf{I} - \Delta \mathbf{K}_\ell \mathbf{H}_{\ell-1})^{-1} + \Delta \mathbf{F}_\ell, \quad \mathbf{J}_{\ell-1} = \mathbf{J}_\ell (\mathbf{I} - \Delta \mathbf{F}_\ell \mathbf{J}_\ell)^{-1} + \Delta \mathbf{K}_\ell. \tag{225}$$

The forward and backward response representatives are

$$\mathbf{R}_\ell = \mathbf{H}_{\ell-1}(\mathbf{I} - \mathbf{J}_{\ell-1}\mathbf{H}_{\ell-1})^{-1}, \quad \mathbf{S}_\ell = \mathbf{J}_\ell (\mathbf{I} - \mathbf{H}_\ell \mathbf{J}_\ell)^{-1}. \tag{226}$$

With Eq. (221), the scalar rows of Eq. (222) give exactly Eq. (220), including for singular history Gram matrices.

Proof. Hold the deterministic coefficients fixed. By Eq. (223), on layers $1, \dots, m$ a change in the supplied $\bar{\mathbf{g}}_m$ gives $\delta \bar{\mathbf{h}}_m = \mathbf{H}_m \delta \bar{\mathbf{g}}_m$. On layers $m+1, \dots, L$, a change in the supplied $\bar{\mathbf{h}}_m$ gives $\delta \bar{\mathbf{g}}_m = \mathbf{J}_m \delta \bar{\mathbf{h}}_m$.

Jacobian recursions. Keep the Gaussian variables fixed. For layers $1, \dots, \ell$, vary the supplied $\bar{\mathbf{g}}_\ell$. For layers $\ell, \dots, L$, vary the supplied $\bar{\mathbf{h}}_{\ell-1}$. Subtracting the corresponding updates in Eq. (222) gives, respectively,

$$\begin{aligned} (\mathbf{I} - \Delta \mathbf{K}_\ell \mathbf{H}_{\ell-1}) \delta \bar{\mathbf{g}}_{\ell-1} &= \delta \bar{\mathbf{g}}_\ell, & \delta \bar{\mathbf{h}}_\ell &= \mathbf{H}_{\ell-1} \delta \bar{\mathbf{g}}_{\ell-1} + \Delta \mathbf{F}_\ell \delta \bar{\mathbf{g}}_\ell, \\ (\mathbf{I} - \Delta \mathbf{F}_\ell \mathbf{J}_\ell) \delta \bar{\mathbf{h}}_\ell &= \delta \bar{\mathbf{h}}_{\ell-1}, & \delta \bar{\mathbf{g}}_{\ell-1} &= \mathbf{J}_\ell \delta \bar{\mathbf{h}}_\ell + \Delta \mathbf{K}_\ell \delta \bar{\mathbf{h}}_{\ell-1}. \end{aligned}$$

Substituting the first equation of each row into the second gives $\mathbf{H}_\ell$ and $\mathbf{J}_{\ell-1}$ in Eq. (225).

Innovation responses. Vary $\boldsymbol{\zeta}_\ell$ alone. The backward update on layers $\ell, \dots, L$ contributes $\sqrt{\Delta}\, \delta \boldsymbol{\zeta}_\ell$ directly to $\delta \bar{\mathbf{g}}_{\ell-1}$, and layers $1, \dots, \ell-1$ give $\delta \bar{\mathbf{h}}_{\ell-1} = \mathbf{H}_{\ell-1} \delta \bar{\mathbf{g}}_{\ell-1}$. For a variation of $\boldsymbol{\xi}_\ell$ alone, the forward update on layers $1, \dots, \ell$ contributes $\sqrt{\Delta}\, \delta \boldsymbol{\xi}_\ell$ directly to $\delta \bar{\mathbf{h}}_\ell$, and layers $\ell+1, \dots, L$ give $\delta \bar{\mathbf{g}}_\ell = \mathbf{J}_\ell \delta \bar{\mathbf{h}}_\ell$. Including the variations of the supplied values gives

$$(\mathbf{I} - \mathbf{J}_{\ell-1}\mathbf{H}_{\ell-1}) \delta \bar{\mathbf{g}}_{\ell-1} = \sqrt{\Delta}\, \delta \boldsymbol{\zeta}_\ell, \qquad (\mathbf{I} - \mathbf{H}_\ell \mathbf{J}_\ell) \delta \bar{\mathbf{h}}_\ell = \sqrt{\Delta}\, \delta \boldsymbol{\xi}_\ell.$$

Multiplying the resulting variations by $\mathbf{H}_{\ell-1}$ and $\mathbf{J}_\ell$, respectively, gives $\mathbf{R}_\ell$ and $\mathbf{S}_\ell$ in Eq. (226). All inverses exist because $(\mathbf{I} - \mathbf{M})^{-1} = \sum_{d=0}^{k} \mathbf{M}^d$ for strictly lower triangular $\mathbf{M}$.

Equality of the response products. Independence of the Gaussian blocks and the response coefficients just computed give

$$\mathbb{E}[\zeta_\ell^{<j} \bar{h}_{\ell-1}^j] = \sqrt{\Delta}\, \bar{C}_{g,\ell}^{<j,<j} (\mathbf{R}_\ell)_{j,<j}^\top, \qquad \mathbb{E}[\xi_\ell^{\leq j} \bar{g}_\ell^j] = \sqrt{\Delta}\, \bar{C}_{h,\ell-1}^{\leq j,<j} (\mathbf{S}_\ell)_{j,<j}^\top.$$

Thus Eq. (177) replaces each representative column r by $C^\dagger C r$, where C is the corresponding history Gram matrix. For the backward response, pad r with zero at index j. For the history X with $C = \mathbb{E}[XX^\top]$, $\mathbb{E}[|X^\top (r - C^\dagger C r)|^2] = 0$. Hence the original and representative response products agree almost surely, including when C is singular.

Finally, at iterate j the blocks on $< j$ are known. The forward pass uses $\bar{h}_{\ell-1}^j$ and $(\mathbf{H}_{\ell-1})_{j,<j}$ to obtain the covariance of ξ_ℓ^j and the row $(\mathbf{F}_\ell)_{j,<j}$, then advances $\bar{h}_\ell^j$ and $(\mathbf{H}_\ell)_{j,<j}$. The backward pass starts at $\bar{g}_L^j = \bar{v}^j$ and uses $\bar{g}_\ell^j$ and $(\mathbf{J}_\ell)_{j,<j}$ to obtain the covariance of ζ_ℓ^j and the row $(\mathbf{K}_\ell)_{j,<j}$, then advances $\bar{g}_{\ell-1}^j$ and $(\mathbf{J}_{\ell-1})_{j,<j}$. Extend the Gaussian blocks by Eq. (183) in each pass. This constructs Eq. (220) in training order. □

F.3.2 Continuous-depth system

Define the covariance kernels by

$$C_h^{s,j}(t) = \mathbb{E}[H_t^s H_t^j], \qquad C_g^{s,j}(t) = \mathbb{E}[G_t^s G_t^j].$$

The response matrices are

$$\mathbf{R}(t) = \mathbf{H}(t)\big(\mathbf{I} - \mathbf{J}(t)\mathbf{H}(t)\big)^{-1}, \qquad \mathbf{S}(t) = \mathbf{J}(t)\big(\mathbf{I} - \mathbf{H}(t)\mathbf{J}(t)\big)^{-1}. \tag{227}$$

For $a < j$, the drift coefficients are

$$(\mathbf{F}(t))_{ja} = (\mathbf{R}(t))_{ja} - \eta_c \chi_\infty^a C_h^{a,j}(t), \qquad (\mathbf{K}(t))_{ja} = (\mathbf{S}(t))_{ja} - \eta_c \chi_\infty^a C_g^{a,j}(t). \tag{228}$$

The entries with $a \geq j$ are zero. The Jacobian equations are

$$\begin{aligned} \mathbf{H}' &= \mathbf{F} + \mathbf{H}\mathbf{K}\mathbf{H}, & (\mathbf{H}(0))_{ja} &= -\eta_c \chi_\infty^a q_x^{a,j}, \\ -\mathbf{J}' &= \mathbf{K} + \mathbf{J}\mathbf{F}\mathbf{J}, & (\mathbf{J}(1))_{ja} &= -\eta_c \chi_\infty^a, \qquad a < j. \end{aligned} \tag{229}$$

Indeed, the exact grid difference equations from Eq. (225) are

$$\begin{aligned} \frac{\mathbf{H}_\ell - \mathbf{H}_{\ell-1}}{\Delta} &= \mathbf{F}_\ell + \mathbf{H}_{\ell-1}\mathbf{K}_\ell(\mathbf{H}_\ell - \Delta \mathbf{F}_\ell), \\ \frac{\mathbf{J}_{\ell-1} - \mathbf{J}_\ell}{\Delta} &= \mathbf{K}_\ell + \mathbf{J}_\ell \mathbf{F}_\ell(\mathbf{J}_{\ell-1} - \Delta \mathbf{K}_\ell). \end{aligned} \tag{230}$$

Specializing Eq. (194) to $\phi(h) = h$ and using the representatives Eq. (227) gives

$$H_t^j = H_0^j + \int_0^t (\mathbf{F}(s))_{j,<j} G_s^{<j} \, \mathrm{d}s + W_t^j,$$

$$G_t^j = V^j + \int_t^1 (\mathbf{K}(s))_{j,<j} H_s^{<j} \, \mathrm{d}s + \widetilde{W}_1^j - \widetilde{W}_t^j, \tag{231}$$

$$H_0^j = \xi_0^j - \eta_c (G_0^{<j})^\top D_{\chi_\infty}^{<j} q_x^{<j,j}, \qquad V^j = v^0 - \eta_c (H_1^{<j})^\top \chi_\infty^{<j}. \tag{232}$$

Here $f_\infty^j = \mathbb{E}[V^j H_1^j]$ and $\chi_\infty^j = \partial_f \mathcal{L}(f_\infty^j, y^j)$ for $j < k$. The Gaussian martingales $W^j, \widetilde{W}^j$ have the law in Eq. (188) with $C_\phi = C_h$. In particular, the backward noise in Eq. (231) is the martingale tail $\widetilde{W}_1^j - \widetilde{W}_t^j$, so BDG from Lemma A.9 applies to this term separately from the drift integral. It is distinct from the nonlinear recursion Eq. (182), where the factor $\phi'(\bar{h}_{\ell-1}^j)$ multiplies each backward noise increment. Applying BDG directly to the resulting weighted sum would require showing that these factors are adapted in backward depth.

F.3.3 Strong depth convergence

Assumption F.7 (Uniform nondegeneracy of older histories)**.** For $k \geq 1$, there are $\kappa > 0$ and $L_0 < \infty$ such that for every $L \geq L_0$,

$$\min_{1 \leq \ell \leq L} \min\left\{ \lambda_{\min}\left(\bar{C}_{h,\ell-1}^{<k,<k,L}\right), \lambda_{\min}\left(\bar{C}_{g,\ell}^{<k,<k,L}\right) \right\} \geq \kappa. \tag{233}$$

At $k = 0$ the condition is vacuous and we take $L_0 = 1$.

Restore the superscript L on grid quantities and let $t_\ell = \ell\Delta$, where $\Delta = L^{-1}$. For $t \in [t_{\ell-1}, t_\ell]$, set $\theta = (t - t_{\ell-1})/\Delta$ and define the affine interpolants of the Jacobians by

$$\mathbf{Q}^L(t) = (1-\theta)\mathbf{Q}_{\ell-1}^L + \theta \mathbf{Q}_\ell^L, \qquad \mathbf{Q} \in \{\mathbf{H}, \mathbf{J}\}.$$

For the response and drift matrices, use the step profiles

$$\mathbf{Q}^L(t) = \mathbf{Q}_\ell^L, \qquad t \in (t_{\ell-1}, t_\ell], \qquad \mathbf{Q} \in \{\mathbf{R}, \mathbf{S}, \mathbf{F}, \mathbf{K}\}.$$

Let $W^{j,L}$ and $\widetilde{W}^{j,L}$ be the coupled Gaussian martingales defined below, with grid increments

$$W_{t_\ell}^{j,L} - W_{t_{\ell-1}}^{j,L} = \sqrt{\Delta}\, \xi_\ell^{j,L}, \qquad \widetilde{W}_{t_\ell}^{j,L} - \widetilde{W}_{t_{\ell-1}}^{j,L} = \sqrt{\Delta}\, \zeta_\ell^{j,L}.$$

For $t \in [t_{\ell-1}, t_\ell]$, define

$$\bar{h}^{j,L}(t) = \bar{h}^{j,L}_{\ell-1} + (t - t_{\ell-1})(\mathbf{F}^L_\ell)_{j,<j}\bar{g}^{<j,L}_\ell + W^{j,L}_t - W^{j,L}_{t_{\ell-1}},$$
$$\bar{g}^{j,L}(t) = \bar{g}^{j,L}_\ell + (t_\ell - t)(\mathbf{K}^L_\ell)_{j,<j}\bar{h}^{<j,L}_{\ell-1} + \widetilde{W}^{j,L}_{t_\ell} - \widetilde{W}^{j,L}_t. \tag{234}$$

At $t = t_\ell$, the first formula gives the forward update in Eq. (222). At $t = t_{\ell-1}$, the second gives the backward update. Thus the interpolants agree with the discrete states at every grid point.

Theorem F.8 (Depth convergence over finitely many updates)**.** *Under the data and loss conditions above and Assumption F.7, the continuous equations Eq. (227) through Eq. (232) determine a unique continuous centered Gaussian law. Under the common ordered Brownian coupling Eq. (186), for every $p \geq 2$ and $L \geq L_0$,*

$$\max_{0\leq j\leq k}\left(\left\|\sup_{0\leq t\leq 1}|\bar{h}^{j,L}(t) - H^j_t|\right\|_{L^p} + \left\|\sup_{0\leq t\leq 1}|\bar{g}^{j,L}(t) - G^j_t|\right\|_{L^p}\right) \leq C_{p,k}L^{-1/2}. \tag{235}$$

The endpoint kernel profiles converge uniformly at the same rate. This rate also holds for the response representatives $\mathbf{R}^L, \mathbf{S}^L$, drift profiles $\mathbf{F}^L, \mathbf{K}^L$, and Jacobian interpolants $\mathbf{H}^L, \mathbf{J}^L$. The readouts $\bar{f}^{j,L}$ converge to f^j_∞ at rate $L^{-1/2}$. The conditional variance of the current innovation at iterate k may vanish.

Proof. Use the same Brownian motions for every grid and for the limit, as in Eq. (186) and Eq. (188). The discrete Brownian integrands on $(t_{\ell-1}, t_\ell]$ are the constant rows of σ^L_ℓ and τ^L_ℓ. The continuous integrands are the rows of $\sigma(s)$ and $\tau(s)$. All four factors are given by Eq. (183), with C_h in place of C_ϕ.

We construct and compare the solutions in the same induction over j. Fix $p \geq 2$. Suppose that the iterates $a < j$ have been constructed with $O(\sqrt{\Delta})$ errors for their paths, Jacobians, kernels, responses, drift coefficients, and losses. The continuous states have $1/2$-Hölder increments in L^p, and their Jacobian rows are Lipschitz, by Eq. (231) and Eq. (229). Hence, for $a < j$ and $s \in (t_{\ell-1}, t_\ell]$,

$$\|\bar{h}^{a,L}_{\ell-1} - H^a_s\|_{L^p} + \|\bar{g}^{a,L}_\ell - G^a_s\|_{L^p} \leq C_p\sqrt{\Delta}.$$

Cauchy–Schwarz gives the same rate for the older Gram blocks. For $j \geq 1$, the grid blocks on $< j$ are principal submatrices of the blocks in Eq. (233). Their limiting blocks therefore satisfy $\lambda_{\min}(C_h^{<j,<j}(s)), \lambda_{\min}(C_g^{<j,<j}(s)) \geq \kappa$, including at $s = 0$ by continuity.

Construct the current iterate. In Eq. (227) and Eq. (229), the current forward Jacobian row $(\mathbf{H})_{j,<j}$ is multiplied only by completed blocks. Also $C_h^{a,j}(s) = \mathbb{E}[H^a_s H^j_s]$ is a bounded linear functional of H^j_s for $a < j$. Thus the forward pair $(H^j, (\mathbf{H})_{j,<j})$ has globally Lipschitz drift and Jacobian terms, with constants bounded over depth by the older iterates. The same holds for the backward pair $(G^j, (\mathbf{J})_{j,<j})$.

For fixed older states, let P be the L^2 projection onto their span. In the ordered diffusion row, the older coordinates come from PX, and the final entry is $\|(\mathbf{I} - P)X\|_{L^2}$ for the current state X. For two candidates X, Y, the squared row difference is at most

$$\|P(X - Y)\|^2_{L^2} + \big|\|(\mathbf{I} - P)X\|_{L^2} - \|(\mathbf{I} - P)Y\|_{L^2}\big|^2 \leq \|X - Y\|^2_{L^2}.$$

At $j = 0$, $P = 0$ and both Jacobian rows are empty. On a depth interval of length u, the integral map for either pair has Lipschitz constant at most $Cu + \sqrt{u}$ in the sum of the supremum L^2 state norm and supremum Jacobian row norm. The drift and Jacobian integrals give Cu, and Itô isometry gives $\sqrt{u}$. The constant is uniform over depth, so successive contractions reach depth 1 for the forward pair and depth 0 for the backward pair. The integral equations give centered Gaussian states with continuous paths and bounded maximal L^p moments. The deterministic kernels determine the joint Gaussian noise law, which can be realized with the ordered drivers. For given coefficients and noises, the state equations determine the paths in training order. The same contraction therefore gives uniqueness in law.

Compare the current iterate with the grid. The boundary equations Eq. (224) and Eq. (232), together with Eq. (229), give $O(\sqrt{\Delta})$ errors in the current state and Jacobian boundary values by the older

path and loss estimates. At $j = 0$ the boundary values agree. Set

$$E_j(t) = \left\| \sup_{0\le r\le t} |\bar{h}^{j,L}(r) - H^j_r| \right\|^2_{L^p} + \sup_{0\le r\le t} |(\mathbf{H}^L(r) - \mathbf{H}(r))_{j,<j}|^2,$$
$$\widetilde{E}_j(t) = \left\| \sup_{t\le r\le 1} |\bar{g}^{j,L}(r) - G^j_r| \right\|^2_{L^p} + \sup_{t\le r\le 1} |(\mathbf{J}^L(r) - \mathbf{J}(r))_{j,<j}|^2.$$

On $s \in (t_{\ell-1}, t_\ell]$, add and subtract the continuous states at $t_{\ell-1}$ for H^j and at t_ℓ for G^j. The increment bounds give

$$\|\bar{h}^{j,L}_{\ell-1} - H^j_s\|_{L^p} \le E_j(s)^{1/2} + C_p\sqrt{\Delta},$$
$$\|\bar{g}^{j,L}_{\ell} - G^j_s\|_{L^p} \le \widetilde{E}_j(s)^{1/2} + C_p\sqrt{\Delta}.$$

The corresponding Jacobian comparisons at these endpoints cost $C\Delta$ in place of $C_p\sqrt{\Delta}$.

To compare the diffusion rows, first change the current state with the older grid states fixed, using the projection bound above. Then change the older states with the current limit state X fixed. For an older column U with Gram matrix C and triangular factor A as in Eq. (183), the two parts of the row are

$$\mathbb{E}[XU^\top]C^{-1}A, \qquad \|X - \mathbb{E}[XU^\top]C^{-1}U\|_{L^2}.$$

The older columns and their Gram blocks differ by $O(\sqrt{\Delta})$. On these bounded blocks with lower bound κ, inversion and Cholesky factorization are Lipschitz, the latter by its scalar recursion with positive pivots. Cauchy–Schwarz and the reverse triangle inequality then give an $O(\sqrt{\Delta})$ difference in both displayed parts. Thus

$$|\sigma^L_{\ell,j,\le j} - \sigma_{j,\le j}(s)| \le \|\bar{h}^{j,L}_{\ell-1} - H^j_s\|_{L^2} + C\sqrt{\Delta},$$
$$|\tau^L_{\ell,j,\le j} - \tau_{j,\le j}(s)| \le \|\bar{g}^{j,L}_{\ell} - G^j_s\|_{L^2} + C\sqrt{\Delta}. \tag{236}$$

For $j = 0$, the scalar diffusion coefficients are L^2 norms, so these bounds follow from the reverse triangle inequality alone.

The response formulas Eq. (226) and the exact Jacobian differences Eq. (230) have multipliers of the current row that are polynomials in the older blocks. Those multipliers differ from their limits by $O(\sqrt{\Delta})$. The cross-kernels in Eq. (221) satisfy the product estimate, for example,

$$|\bar{C}^{a,j,L}_{h,\ell-1} - C^{a,j}_h(s)| \le \|\bar{h}^{a,L}_{\ell-1}\|_{L^2}\|\bar{h}^{j,L}_{\ell-1} - H^j_s\|_{L^2} + \|\bar{h}^{a,L}_{\ell-1} - H^a_s\|_{L^2}\|H^j_s\|_{L^2}.$$

Combining the endpoint and loss estimates with Eq. (236) yields the analogue of Eq. (161):

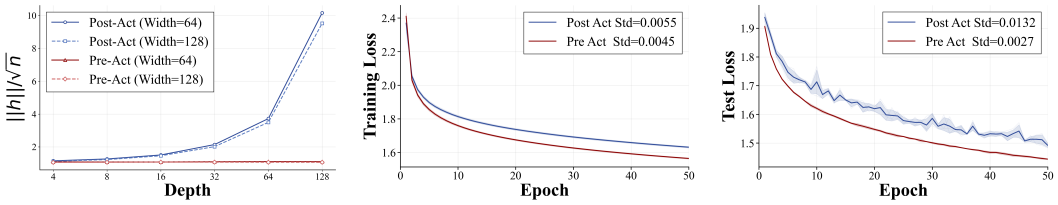
$$\begin{aligned} &|\sigma^L_{\ell,j,\le j} - \sigma_{j,\le j}(s)|^2 + |(\mathbf{F}^L_\ell - \mathbf{F}(s))_{j,<j}|^2 + |((\mathbf{H}^L)'(s) - \mathbf{H}'(s))_{j,<j}|^2 \\ &\qquad \le C_p(E_j(s) + \Delta), \\ &|\tau^L_{\ell,j,\le j} - \tau_{j,\le j}(s)|^2 + |(\mathbf{K}^L_\ell - \mathbf{K}(s))_{j,<j}|^2 + |((\mathbf{J}^L)'(s) - \mathbf{J}'(s))_{j,<j}|^2 \\ &\qquad \le C_p(\widetilde{E}_j(s) + \Delta). \end{aligned} \tag{237}$$

Subtract Eq. (231) from Eq. (234), summing over the cells. The forward drift difference on a cell is

$$(\mathbf{F}^L_\ell - \mathbf{F}(s))_{j,<j}\bar{g}^{<j,L}_\ell + (\mathbf{F}(s))_{j,<j}(\bar{g}^{<j,L}_\ell - G^{<j}_s).$$

Its squared L^p norm is bounded by $C_p(E_j(s) + \Delta)$. For the backward drift, use $\mathbf{K}$ in place of $\mathbf{F}$ and compare $\bar{h}^{<j,L}_{\ell-1}$ with $H^{<j}_s$. This gives the bound $C_p(\widetilde{E}_j(s) + \Delta)$. For the noise, integrate the squared row differences over the cells and apply BDG (Lemma A.9). This gives

$$\left\| \sup_{0\le r\le t} |W^{j,L}_r - W^j_r| \right\|^2_{L^p} \le C_p \int_0^t (E_j(s) + \Delta)\,\mathrm{d}s,$$
$$\left\| \sup_{t\le r\le 1} |(\widetilde{W}^{j,L}_1 - \widetilde{W}^{j,L}_r) - (\widetilde{W}^j_1 - \widetilde{W}^j_r)| \right\|^2_{L^p} \le C_p \int_t^1 (\widetilde{E}_j(s) + \Delta)\,\mathrm{d}s. \tag{238}$$



(a) Feature norms vs. depth **(b) Training loss** **(c) Test loss**

Figure 4: **Pre-activation and post-activation ResNets under depth-μP.** Panel (a) compares feature norms across depths. Panels (b) and (c) show training and test losses for depth-64, width-128 ReLU ResNets trained on CIFAR-10 with mini-batch SGD and batch size 128.

The backward estimate uses the difference of two increments starting at t. Integrating the drift and Jacobian errors by Cauchy–Schwarz in depth, and adding the boundary errors, now gives

$$E_j(t) \le C_p\Delta + C_p \int_0^t E_j(s)\,\mathrm{d}s, \qquad \widetilde{E}_j(t) \le C_p\Delta + C_p \int_t^1 \widetilde{E}_j(s)\,\mathrm{d}s.$$

Gronwall yields $\sup_t(E_j(t) + \widetilde{E}_j(t)) \le C_p\Delta$. The product estimate gives the same $O(\sqrt{\Delta})$ rate for the endpoint kernels and the readout $\bar{f}^{j,L} = \mathbb{E}[\bar{g}_L^{j,L}\bar{h}_L^{j,L}]$. The Lipschitz loss derivative gives the rate for $\bar{\chi}^{j,L}$. The response and drift estimates follow from the polynomial response formulas and Eq. (221), completing the induction. □

Remark F.9 (Scope of the Gram lower bound). Condition Eq. (233) controls only the older Gram blocks used in the ordered coupling Eq. (186). The residual estimate in Eq. (236) uses a norm, so it also covers a zero residual at the current iterate. The fixed-depth response formulas in Lemma F.6 also hold for singular history Gram matrices. For the depth limit, κ in Eq. (233) is uniform in L. The width hypothesis in Proposition F.4 is stated at each fixed L.

F.3.4 Iterated width–depth limit

Corollary F.10. *Suppose Assumption F.7 holds and the fixed-depth hypotheses of Proposition F.4 hold for each $L \ge L_0$. For every $p \ge 2$, there is a coupling such that*

$$\max_{\substack{j\le k\\ \ell\le L}} \sup_{1\le i\le n} \left(\|h^j_{\ell,i} - H^j_{t_\ell,i}\|_{L^p} + \|g^j_{\ell,i} - G^j_{t_\ell,i}\|_{L^p}\right) + \max_{j\le k}\|f^j - f^j_\infty\|_{L^p}$$
$$\le C_{p,L,k}n^{-1/2} + C_{p,k}L^{-1/2},$$

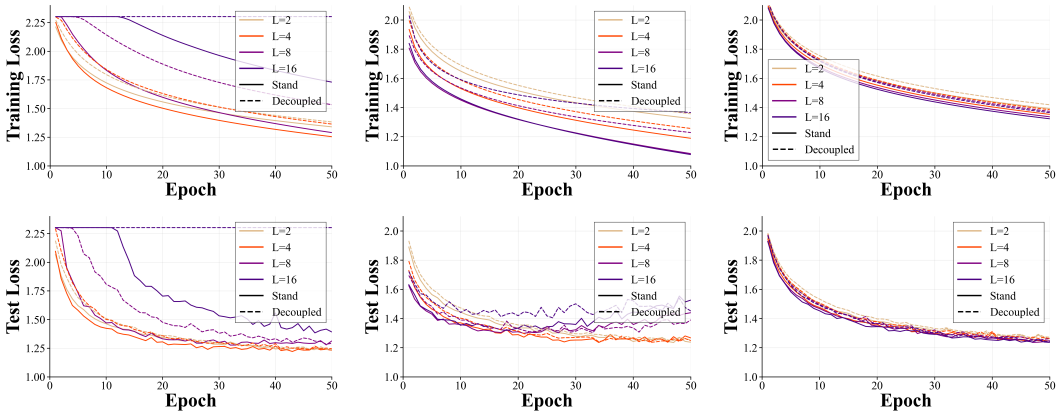
where (H_i, G_i) are independent copies of the joint continuous-depth system Eq. (231). The limit is taken as $n \to \infty$ followed by $L \to \infty$.

Proof. The width and depth couplings have the same joint fixed-depth mean-field marginal $(\bar{h}^{j,L}, \bar{g}^{j,L})_{j\le k}$ on the grid by Lemma F.6. Join the couplings by sampling their conditional laws given that marginal. Apply the triangle inequality, Proposition F.4, and Theorem F.8 to the states and readouts. □

G Additional Experiments

In this section, we provide supplementary experiments that complement the results reported in the main paper. Unless otherwise specified, our default setup trains the ResNet in Eq. (1) using mini-batch SGD with batch size 128 and effective learning rate $\eta_c = 0.01$, across widths $n \in \{128, 256\}$ and depths $L \in \{2, 4, 8, 16\}$.

Pre- vs. Post-activation ResNets. We provide additional experiments supporting the architectural choice used in the main paper. In particular, we compare pre-activation and post-activation ResNets under depth-μP scaling. As suggested by Proposition 3.1, the pre-activation design used in our theoretical analysis is more stable at large depth, whereas post-activation ResNets can exhibit rapid feature-norm growth when the activation satisfies the positive dominance condition. Figure 4 supports this behavior and shows that pre-activation also leads to more stable train/test loss curves.



(a) Vanilla DNN **(b) μP ResNet** **(c) Depth-μP ResNet**

Figure 5: **Loss curves with standard and decoupled backpropagation at width 256.** The top row shows training loss and the bottom row shows test loss. The three columns correspond to panels (a)–(c) of Figure 2.

Standard and decoupled backpropagation at width 256. Figure 5 repeats the first three comparisons in Figure 2 at width 256. In the standard runs, backpropagation reuses the forward weights. In the decoupled runs, the backward weights start from an independent copy of the initial forward weights and receive the same subsequent updates as the forward weights within that run. Compared with width 128, the vanilla DNN and μP ResNet runs are more stable, although visible gaps between the standard and decoupled loss curves remain. The depth-μP ResNets continue to show close agreement in both training and test loss at width 256. This agreement in loss is compatible with the one-update analysis, in which the response induced by weight reuse can persist in the continuous-depth feature dynamics.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main theoretical claims: reused-weight forward–backward coupling is a finite-width effect at initialization, but can survive the infinite-width limit during training, and is suppressed as a higher-order depth effect under depth-μP scaling. The paper also specifies the scope of the results, namely one-layer ResNets, stream SGD, fixed finite training horizon, and nondegeneracy assumptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

Justification: The paper includes a limitations discussion in the Conclusion section, stating that the analysis focuses on one-layer ResNets under depth-μP scaling trained by stream SGD, assumes nondegeneracy conditions, and proves convergence for a fixed finite training horizon. It also notes that extensions to Transformers, adaptive optimizers, longer training horizons, and optimization/generalization guarantees remain open.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The theoretical statements explicitly state their assumptions, including Lipschitz/regularity assumptions, nondegeneracy conditions, fixed finite training horizon, and the sequential width-depth limiting regime. Proofs are provided in the appendix, including Gaussian representation arguments, Tensor Program width-limit derivations, and depth-convergence estimates. When using Tensor Programs, we cite the relevant Master Theorem and explain how it applies to the finite-horizon SGD computation graph.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper reports the model class, dataset, optimizer, width/depth settings, and experimental protocols used to illustrate the theoretical predictions, and provides reproducible code in the supplementary materials. The experiments support and visualize the predicted scaling behavior rather than establish a new empirical benchmark.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The experiments use standard publicly available datasets, such as CIFAR-10, and the code to reproduce the experimental results is provided in the supplementary material, along with implementation details.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the experimental setting, including the use of SGD, stream-style training where relevant, depth and width variations, and the quantities measured to compare finite-network behavior with the predicted limiting dynamics. Additional implementation details are provided in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments are mainly illustrative. Where variability is relevant, the paper reports results across multiple random initialization seeds and describes the variability observed across these runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This paper mainly verifies the theoretical derivation results through experiments, without involving large-scale GPU calculations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: The work is theoretical and uses standard public datasets only for illustrative experiments. It does not involve human subjects, private data, high-risk model release, or applications that raise special ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theoretical research on feature-learning dynamics in neural networks and does not introduce a deployed system, dataset, or application-specific method. Any societal impact would be indirect through future improvements in understanding and designing deep learning systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper solely utilizes the standard open-source CIFAR-10 dataset for theoretical validation. and does not involve any high-risk models or data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used the publicly available CIFAR-10 dataset for our experiments. Its usage is explicitly mentioned in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We submitted detailed code in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for writing and editing purposes to improve the language of the manuscript, and do not constitute an important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.