Understanding Scaling Laws via Neural Feature Learning Dynamics

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

Recently, deep neural networks have revolutionized various domains, primarily due to their ability to consistently improve performance when scaling up resources, including model size, data, and compute, a phenomenon formalized as scaling laws. Yet, the theoretical basis of these principles remains unclear: why scaling works and when it breaks down. We address this gap by analyzing the feature learning dynamics of ResNets trained with SGD. In the joint infinite-width–depth limit, we show that feature evolution is governed by a coupled forward–backward stochastic system, which we term the *neural feature learning dynamic system*. This framework clarifies the mechanisms underlying scaling laws and offers a new mathematical tool for studying deep learning dynamics.

1 Introduction

In recent years, deep neural networks (DNNs) have achieved remarkable success across diverse domains. A key empirical observation behind this success is that performance continuously improves as resources—including model size, data, and compute—are scaled up, a phenomenon formalized as *scaling laws* [19, 16]. These principles have guided the development of many state-of-the-art large-scale models, often with hundreds of billions of parameters, including large language models (LLMs) [4, 7, 29], vision transformers (ViTs) [8], and deep generative models [15, 25, 27]. However, training these massive models is not without challenges. As models grow larger, they often encounter severe training instabilities—such as loss spikes and exploding gradients [20, 6]. Furthermore, they also exhibit diminishing returns in test performance, especially when data does not scale proportionally [19, 16]. These contrasting outcomes—breakthroughs on one hand and breakdowns on the other—reveal a fundamental gap in our theoretical understanding: *Why does scaling up consistently improve performance, yet still collapse in some cases?*

To approach this question, one widely studied framework is the Neural Tangent Kernel (NTK) theory. In the infinite-width limit, DNNs trained with gradient-based methods evolve linearly around their initialization, becoming kernel machines governed by a fixed NTK [18]. This connection helps explain why sufficiently wide networks can perfectly fit training data while still generalizing well [1, 22]. However, NTK theory captures only the so-called *lazy training regime* [5, 31], where parameter changes are limited and features remain largely fixed. Consequently, the NTK framework cannot account for the rich representation learning that underlies key paradigms of modern deep learning, such as parameter-efficient fine-tuning [17], in-context learning [4, 9], and chain-of-thought reasoning [30].

In contrast, a complementary line of work studies *feature learning (FL) regimes*, where neural networks actively learn features even at scale. The *maximal update parameterization (μP)* framework

[*] Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

[34] introduces a general scaling rule that preserves nontrivial FL as network width grows. Notably, μP enables hyperparameters tuned on small models to transfer reliably to larger ones [35], a critical advantage for large-scale training. Despite these advances, most existing analyses in the FL regime are restricted to shallow networks and thus cannot fully capture how features evolve in deep architectures. This limitation is important because the success of deep learning is widely attributed to *depth* [2, 21]. While recent work has begun exploring the role of depth [36, 3], our theoretical understanding of how depth shapes feature learning dynamics in ultra-deep networks remains limited.

To address the gap in understanding scaling laws, we extend the FL perspective by analyzing feature learning dynamics in the joint infinite-width and infinite-depth limit (henceforth referred to as the *joint limit*). Our analysis focuses on randomly initialized residual networks (ResNets) trained with stochastic gradient descent (SGD), a dominant architecture in large-scale models [13, 8, 28]. In this limit, we show that forward feature propagation in ResNets is described by a *stochastic differential equation (SDE)*, where depth plays the role of a continuous time variable; during training, backpropagation induces a feedback process—modeled as a *backward SDE*—that continuously reshapes the forward SDE by modifying its drift and diffusion terms.

This SDE-based perspective provides a principled understanding of the mechanisms behind scaling laws: scaling succeeds when finite networks closely approximate the SDE limit; it fails when convergence to this limit breaks down. Additionally, the coupled forward–backward SDE system, referred to as the *Feature Learning Dynamics System (FLDS)*, also provides a mathematical foundation for future theoretical advances in studying training behavior and scaling effects beyond the lazy regime. Our contributions are summarized as follows:

- We first analyze ResNets in the joint limit at initialization and prove that the *pre-activation* design is inherently more stable than the post-activation design,[2] as the latter can diverge with certain activation functions (e.g., ReLU).
- Under the pre-activation design, we show that as width and depth both tend to infinity, feature propagation in ResNets converges to a forward SDE. This convergence holds regardless of the relative scaling rates of width and depth, establishing their *commutativity*.
- During training under μP scaling, we show that the forward SDE is dynamically reshaped by a backward SDE induced by backpropagation. Together, these coupled stochastic processes form the *Feature Learning Dynamics System (FLDS)*.
- We revisit the *gradient independence assumption (GIA)*—commonly valid at initialization in the infinite-width limit, where forward weights $\boldsymbol{W}$ and backward weights $\boldsymbol{W}^\top$ are treated as independent. We show that this assumption, which generally fails during training at finite depth, becomes valid again in the infinite-depth limit.

2 Preliminaries

Depth-adapted ResNets. Give $\boldsymbol{x} \in \mathbb{R}^d$, we consider a ResNet defined through a residual stream:

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

where ϕ is an activation function, $\boldsymbol{U} \in \mathbb{R}^{n\times d}$, $\boldsymbol{W}_\ell \in \mathbb{R}^{n\times n}$, and $\boldsymbol{v} \in \mathbb{R}^n$ are trainable parameters. These parameters, denoted collectively by $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{U}, \{\boldsymbol{W}_\ell\}, \boldsymbol{v})$, are randomly initialized [10, 12]: $\boldsymbol{v}_i, \boldsymbol{W}_{\ell,ij}, \boldsymbol{U}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

Given a loss function $\mathcal{L}$, we train the ResNet f via gradient descent (GD) $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}(f^{(k)}, y^{(k)})$, where $f^{(k)} := f(\boldsymbol{x}^{(k)};\boldsymbol{\theta}^{(k)})$ is evaluated on a single data point $(\boldsymbol{x}^{(k)}, y^{(k)})$ sampled at iteration k by following the standard in the Tensor Program (TP) literature [34, 35, 36] to simplify the analysis in the mean-field dynamics.

Tensor Programs and μP Parameterization. The TP framework [32, 34] provides a unified language for expressing the forward and backward computations of neural networks and analyzing their infinite-width limits. Formally, a TP is a sequence of vectors recursively computed from an initial

[2] Unless additional corrective operations are applied (see, *e.g.*, [36]).

set of random parameters using basic TP operations (*e.g.*, MatMul, Nonlin, Moment). Importantly, TP remains valid during training, when the recomputed vectors are expressed via valid TP operations.

A central theoretical result of TP is the *Master Theorem* [34, Theorem 7.4], capturing the mean-field behavior of program variables. Specifically, for a finite set of program variables $\{\boldsymbol{h}_s\}_{s=1}^M \in \mathbb{R}^n$ and a sufficiently regular function $\psi : \mathbb{R}^M \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\boldsymbol{h}_{1,i}, \dots, \boldsymbol{h}_{M,i}) \overset{a.s.}{\to} \mathbb{E}\big[\psi(Z^{\boldsymbol{h}_1}, \dots, Z^{\boldsymbol{h}_M})\big], \quad \text{as } n \to \infty, \tag{2}$$

where $Z^{\boldsymbol{h}_s}$ denotes the *mean-field limit* of $\boldsymbol{h}_s$. This characterization enables a rigorous description of signal propagation and training dynamics in infinitely wide networks. Building on the Master Theorem, the μP [34] was shown to maximize feature evolution per gradient update and enables hyperparameter transfer from small to large models [35]. Our setup, defined in Eq. (1), also lies in the μP regime.

3 Main Results: Neural Feature Learning Dynamics System

The Pre- vs. Post-Activation Debate It is important to first distinguish between two common ResNet designs: the *pre-activation* style, as defined in Eq. (1), and the *post-activation* style, where the skip connection is applied after the activation. Although the post-activation design was first introduced [13], the pre-activation variant [14] has become the preferred choice in modern deep networks [24, 4]. However, this preference has been guided more by practice than theory. The following result provides a theoretical justification for the pre-activation preference when network depth increases.

Proposition 1. *Let ϕ satisfy the following* ***positive dominance*** *condition: there exist nonnegative constants c_1, c_2, not both zero, such that $\mathbb{E}[\phi(xZ)] \geq c_1|x| + c_2$, $\forall x \in \mathbb{R}$, where Z is the standard Gaussian random variable. Then, in a post-activation ResNet, the expected hidden state satisfies:*

$$\mathbb{E}[\boldsymbol{h}_{L,i}] \geq c_1 \left(1 + c_1\sqrt{T/Ln}\right)^L \|\boldsymbol{x}\|/\sqrt{d} + c_2\sqrt{TL}, \quad \forall i \in [n]. \tag{3}$$

This result implies that with common activations such as ReLU (where $c_1 = 1$ and $c_2 = 0$), the hidden states $\boldsymbol{h}_\ell$ can diverge as depth grows—even under stabilizing depth scaling factors $\sqrt{T/L}$.

Neural Feature Propagation as an SDE Given the superior stability, the subsequent analysis focuses on the pre-activation design. We use the synchronous coupling method to prove that, in the joint limit, each coordinate of the hidden state $\boldsymbol{h}_\ell$ can be viewed as a particle whose mean-field dynamics are characterized by a forward SDE.

Proposition 2. *Suppose ϕ is K_1-Lipschitz. In the joint limit $\min(n, L) \to \infty$, the features $\boldsymbol{h}_\ell$ converge to h_t solving the McKean–Vlasov SDE $dh_t = \sigma_t\, dw_t$ with $h_0 \sim \mathcal{N}(0, \|\boldsymbol{x}\|^2/d)$, $\sigma_t^2 = \mathbb{E}[\phi^2(h_t)]$, and $\{w_t\}$ as a standard Brownian motion for $t \in [0, T]$. The convergence holds in mean square with rate $\mathbb{E}|\boldsymbol{h}_{\ell,i} - h_{t_\ell}|^2 \leq C(L^{-1} + n^{-1})$, where $t_\ell = \ell T/L$.*

The bound further shows that the limit is obtained regardless of the relative scaling rates of width and depth. Remarkably, this *commutability* is not universal across network structures [23, 11], highlighting the robustness of the SDE limit for stable hyperparameter transfer across model scales.

Backward SDE Induced by Backpropagation During training, the gradients are computed through backpropagation. To analyze this process in the joint limit, our analysis focuses on the following backward information flow:

$$\boldsymbol{g}_L = n\frac{\partial f}{\partial \boldsymbol{h}_L} = \boldsymbol{v}, \qquad \boldsymbol{g}_{\ell-1} = n\frac{\partial f}{\partial \boldsymbol{h}_\ell} = \boldsymbol{g}_\ell + \sqrt{\frac{T}{Ln}}\phi'(\boldsymbol{h}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell, \quad \forall \ell \in [L], \tag{4}$$

where $\odot$ is element-wise multiplication. Under mild regularity assumptions on ϕ and ϕ', the limiting behaviors of $\boldsymbol{g}_\ell$ in the joint limit are described by a *backward* SDE[3].

[3] Here instead of defining processes with a reverse of time, the term is used to mean that the SDE starts from time T and goes backward to time 0. This is not to be confused with the classic terminology BSDE.

Proposition 3. *Suppose ϕ and ϕ' are K_1- and K_2-Lipschitz. As $\min(n, L) \to \infty$, the gradients $\boldsymbol{g}_\ell$ converge to g_t solving the McKean–Vlasov SDE $dg_t = \tilde{\sigma}_t\, d\tilde{w}_t$ with $g_T \sim \mathcal{N}(0, 1)$, $\tilde{\sigma}_t^2 = |\phi'(h_t)|^2\, \mathbb{E}[g_t]^2$, and $\{\tilde{w}_t\}$ a Brownian motion independent of $\{w_t\}$ in Propostion 2. The convergence holds in mean square with rate $\mathbb{E}|\boldsymbol{g}_{\ell,i} - g_{t_\ell}|^2 \leq C(L^{-1} + n^{-1})$.*

Similar to the forward propagation in Proposition 2, this convergence is also *commutable*. Moreover, the forward and backward SDEs are driven by *independent* Brownian motions. Although Proposition 2–3 study the information propagation at initialization, we will show below that it also holds valid in the infinite-depth limit.

Feature Learning Dynamics Formed by SGD We now investigate how representations evolve during training in the joint limit. We demonstrate here that the SDE view derived in Section 3 continues to hold, and is continuously reshaped by the backward SDE throughout training.

To ensure that the network operates in the FL regime (rather than NTK with asymptotically frozen features), we adopt the μP scaling from the TP framework by using a learning rate $\eta = \eta_c n$, where $\eta_c > 0$ is a constant.

We can inductively show that all $\{\boldsymbol{h}_\ell^{(k)}, \boldsymbol{g}_\ell^{(k)}\}$ are valid TP variables. This enables the application of the Master Theorem [34, Theorem 7.4] to study the training trajectory in the mean-field limit.

Proposition 4. *Suppose $\mathcal{L}'$, ϕ, and ϕ' are pseudo-Lipschitz continuous. Then, as $n \to \infty$, the output $f^{(k)}$ converges a.s. to $\mathring{f}^{(k)} = \mathbb{E}[Z^{\boldsymbol{g}_L^{(k)}} Z^{\boldsymbol{h}_L^{(k)}}]$, where the hidden states evolve recursively as:*

$$Z^{\boldsymbol{h}_\ell^{(k)}} = Z^{\boldsymbol{h}_{\ell-1}^{(k)}} + \sqrt{\tau} Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}^{(k)}} - \tau \sum_{i=0}^{k-1} \eta_c \mathcal{L}'(\mathring{f}^{(i)}, y^{(i)}) \mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}^{(i)}}) \phi(Z^{\boldsymbol{h}_{\ell-1}^{(k)}})] Z^{\boldsymbol{g}_\ell^{(i)}} \tag{5}$$

$$- \tau^2 \sum_{i=0}^{k-1} \eta_c \mathcal{L}'(\mathring{f}^{(i)}, y^{(i)}) \mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}^{(i)}}) \phi(Z^{\boldsymbol{h}_{\ell-1}^{(k)}})] \mathbb{E}[\phi'(Z^{\boldsymbol{h}_{\ell-1}^{(i)}}) \phi'(Z^{\boldsymbol{h}_{\ell-1}^{(k)}})] Z^{\boldsymbol{g}_\ell^{(i)}}, \tag{6}$$

where $\tau = T/L$ and $\{Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}^{(i)}}\}_{\ell,i}$ are centered joint Gaussian whose variance are computed based on [34, Definition 7.3].

The final term in Eq. (6) reflects additional interactions between the forward and backward paths by using the same weights $\boldsymbol{W}_\ell$. From the perspective of the Euler–Maruyama scheme, however, this correlation term vanishes as $\tau \to 0$ (equivalently $L \to \infty$). Thus, in the FL regime, the evolution of feature learning driven by the backpropagation converges to a coupled stochastic system in the joint limit—which we refer to as the *Feature Learning Dynamics System*.

Definition 1 (Feature Learning Dynamics System)**.** *The Feature Learning Dynamics System (FLDS) is a coupled forward–backward SDE system that describes the evolution of hidden states and gradients over training iteration k in the joint limit:*

$$dh_t^{(k)} = -\sum_{i=0}^{k-1} \eta_c \mathcal{L}'(\mathring{f}^{(i)}, y^{(i)}) \mathbb{E}[\phi(h_t^{(i)}) \phi(h_t^{(k)})] g_t^{(i)}\, dt + dw_t^{(k)}, \tag{7}$$

$$dg_t^{(k)} = -\sum_{i=0}^{k-1} \eta_c \mathcal{L}'(\mathring{f}^{(i)}, y^{(i)}) \mathbb{E}[g_t^{(i)} g_t^{(k)}] \phi(h_t^{(i)}) \phi'(h_t^{(k)})\, dt + \phi'(h_t^{(k)})\, d\tilde{w}_t^{(k)}, \tag{8}$$

where the Brownian motions $\{w_t^{(k)}\}_k$ and $\{\tilde{w}_t^{(k)}\}_k$ have time-varying covariance:

$$\frac{d}{dt}\mathbb{E}[w_t^{(i)} w_t^{(j)}] = \boldsymbol{\Sigma}_{t,ij} := \mathbb{E}[\phi(h_t^{(i)}) \phi(h_t^{(j)})], \quad \frac{d}{dt}\mathbb{E}[\tilde{w}_t^{(i)} \tilde{w}_t^{(j)}] = \Theta_{t,ij} := \mathbb{E}[g_t^{(i)} g_t^{(j)}]. \tag{9}$$

Theorem 1. *Suppose (i) $\mathcal{L}'$, ϕ, and ϕ' are Lipschitz continuous; and (ii) there exists a solution to the FLDS system such that $\{\boldsymbol{\Sigma}_{t,k\times k}, \Theta_{t,k\times k}\}_{t\in[0,T]}$ are uniformly strictly positive definite for each k. As $n \to \infty$ followed by $L \to \infty$, the ResNet output $f^{(k)}$ converges a.s. to $\mathring{f}^{(k)} = \mathbb{E}[g_T^{(k)} h_T^{(k)}]$, where $h_t^{(k)}$ and $g_t^{(k)}$ evolve under the FLDS in Definition 1.*

This limiting stochastic system is fundamentally different from NTK dynamics. Rather than asymptotically freezing representations at initialization, the system continuously refines features via gradient feedback. Moreover, the noise covariance evolves over time, reflecting the accumulation of training effects. These dynamics capture the essence of feature learning in deep networks and open new research directions for understanding large-scale deep learning from a mean-field perspective.

Acknowledgment R. Wu was supported in part by NSF DMS-2308120 and Simons Foundation travel grant MP-TSM-00002346

References

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruoqi Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] 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.

[4] 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.

[5] Lena Chizat and Francis Bach. On the lazy training of neural networks. In *Advances in Neural Information Processing Systems*, 2019.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 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] Tianyu Gao, Xuan Yao, and Danqi Chen. Rethinking the role of demonstration in in-context learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

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

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[13] 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*, pages 770–778, 2016.

[14] 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*, pages 630–645. Springer, 2016.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Da Cai, Trevor Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Wehrmeister, Julia Bethge, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 2022.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Callanan, Prafulla Dhariwal, Reza Ghasemipour, Tom Henighan, Geoffrey Ho, Hyeon-Woo Jun, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[20] Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, et al. What language model to train if you have one million gpu hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, 2022.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] Jaehoon Lee, Jascha Sohl-Dickstein, Kyunghyun Cho, Jeffrey Pennington, Reza Alizadeh, Dumitru Erhan, and Sergey Levine. Wide neural networks of any depth evolve as linear models under gradient descent. In *International Conference on Learning Representations (ICLR)*, 2019.

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

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[26] A-S. Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.

[27] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Edouard Grave, Guillaume Lample, , et al. LLaMA: Open and efficient foundation language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[30] Jason Wei, Yi Tay, Rishi Bommasani, Da Dong, Yi Huang, Denny Zhou, Peng Yang, Guang Ma, Hao Zhou, Jian Dong, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[31] Blake Woodworth, Amir Ghorbani, Yi Li, Tengyu Ma, and Yura H. Al-Saedi. The wide regime of neural networks: A deep learning perspective. In *Advances in Neural Information Processing Systems*, 2020.

[32] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

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

[34] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021.

[35] 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 A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[36] 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.

A Useful Mathematical Results

Lemma 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 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

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

B Analysis of Post-Activation Design — Proof of Proposition 1

Recall that ϕ is assumed to be positive dominate in the statement of Proposition 1, that is, there exist nonnegative 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. Now consider a post-activation ResNet $\{\boldsymbol{h}_\ell\}$ given by

$$\begin{aligned}
\boldsymbol{h}_0 =& \phi(\boldsymbol{U}\boldsymbol{x}), \\
\boldsymbol{h}_\ell =& \boldsymbol{h}_{\ell-1} + \sqrt{\frac{T}{L}}\phi(\boldsymbol{W}_\ell \boldsymbol{h}_{\ell-1}), \quad \forall \ell \in \{1, 2, \cdots, L\},
\end{aligned}$$

where $\boldsymbol{W}_{\ell,ij}$ are independent $\mathcal{N}(0, \frac{1}{n})$ and $\boldsymbol{U}_{ij}$ are independent $\mathcal{N}(0, \frac{1}{d})$.

First note that $(\boldsymbol{U}\boldsymbol{x})_i \sim \mathcal{N}(0, \frac{\|\boldsymbol{x}\|^2}{d})$ and hence $\mathbb{E}[\boldsymbol{h}_{0,i}] \geq c_1 \frac{\|\boldsymbol{x}\|}{\sqrt{d}} + c_2$. Let $\mathcal{B}_\ell$ be the σ-algebra generated by $\{\boldsymbol{h}_0, \cdots, \boldsymbol{h}_{\ell-1}\}$. Then, observe that

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

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

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

Therefore, we obtain

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

Therefore we obtain

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

as $L \to \infty$, provided either $c_1 > 0$ or $c_2 > 0$. Hence, $\mathbb{E}[\|\boldsymbol{h}_L\|] \geq \mathbb{E}[\boldsymbol{h}_{L,i}] \to \infty$.

C Width-first convergence — Proofs of Propositions 2 and 3

C.1 First forward — Proof of Proposition 2

Suppose ϕ satisfies the assumption in Proposition 2 throughout. Consider

$$\boldsymbol{h}_\ell = \boldsymbol{h}_{\ell-1} + \sqrt{\frac{T}{Ln}}\boldsymbol{W}_\ell \phi(\boldsymbol{h}_{\ell-1})$$

which is equivalent to

$$\boldsymbol{h}_\ell = \boldsymbol{h}_{\ell-1} + \sqrt{\frac{T}{Ln}}\|\phi(\boldsymbol{h}_\ell)\|\boldsymbol{z}_\ell, \quad \boldsymbol{z}_\ell \sim \mathcal{N}(0, \boldsymbol{I}_n).$$

To distinguish from quantities after taking limits of $n \to \infty$ and $L \to \infty$, we add superscripts and write each coordinate as

$$h^{n,L}_{\ell,i} = h^{n,L}_{\ell-1,i} + \sqrt{\frac{T}{Ln}}\|\phi(\boldsymbol{h}^{n,L}_{\ell-1})\| z_{\ell,i}.$$

We want to show that each coordinate converges to

$$h^L_{\ell,i} = h^L_{\ell-1,i} + \sqrt{\frac{T}{L}}\sqrt{\mathbb{E}\phi^2(h^L_{\ell-1,i})} z_{\ell,i}$$

as $n \to \infty$.

Lemma 3. *For each $i \in \mathbb{N}$, $\sup_{L\geq 1}\sup_{\ell=0,\dots,L} \mathbb{E}[h^L_{\ell,i}]^4 < \infty$ and $\inf_{L\geq 1}\inf_{\ell=0,\dots,L} \mathbb{E}\phi^2(h^L_{\ell,i}) > 0$.*

Proof of Lemma 3. By symmetry, we omit the subscript i. Using independence of z_ℓ, we have

$$\begin{aligned}\mathbb{E}[h^L_\ell]^4 &\leq C\mathbb{E}[h^L_0]^4 + C\mathbb{E}\left[\sum_{u=1}^{\ell}\sqrt{\frac{T}{L}}\sqrt{\mathbb{E}\phi^2(h^L_{u-1})}z_u\right]^4 \\ &\leq C + \frac{C}{L}\sum_{u=1}^{\ell}\mathbb{E}\phi^4(h^L_{u-1}) \leq C + \frac{C_{K_1}}{L}\sum_{u=0}^{\ell-1}\mathbb{E}[h^L_u]^4.\end{aligned}$$

It then follows from discrete Gronwall's inequality (Lemma 2) that

$$\mathbb{E}[h^L_\ell]^4 \leq Ce^{C_{K_1}\ell/L}. \tag{10}$$

This gives the first assertion.

For the second assertion, note that ϕ is a continuous function and not identically zero. So there exists some interval $(a,b) \subset \mathbb{R}$ such that $\inf_{a<x<b}\phi^2(x) > 0$. Since $h^L_{\ell-1}$ and z_ℓ are independent, we have

$$\text{Var}(h^L_\ell) \geq \text{Var}(h^L_{\ell-1}) \geq \cdots \geq \text{Var}(h^L_0) = C_1 > 0.$$

Also

$$\text{Var}(h^L_\ell) \leq \mathbb{E}[h^L_\ell]^2 \leq C_2.$$

So $\{h^L_\ell\}$ are Gaussian random variables with mean zero and variance in $[C_1, C_2]$. Therefore,

$$\inf_{L\geq 1}\inf_{\ell=0,\dots,L}\mathbb{P}(h^L_\ell \in (a,b)) > 0.$$

This gives the second assertion. □

Proposition 5. *For each $i \in \mathbb{N}$,*

$$\sup_{L\geq 1}\sup_{\ell=0,\dots,L}\mathbb{E}(h^{n,L}_{\ell,i}-h^{L}_{\ell,i})^2 \leq C/n.$$

Proof of Proposition 5. Since $h^{n,L}_{0,i}=h^{L}_{0,i}$, we have

$$\begin{aligned}\mathbb{E}(h^{n,L}_{\ell,i}-h^{L}_{\ell,i})^2 &= \mathbb{E}\left[\sum_{u=1}^{\ell}\left(\frac{\|\phi(h^{n,L}_{u-1})\|}{\sqrt{n}}-\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}\right)\sqrt{\frac{T}{L}}z_{u,i}\right]^2\\ &= \frac{C}{L}\mathbb{E}\sum_{u=1}^{\ell}\left(\frac{\|\phi(h^{n,L}_{u-1})\|}{\sqrt{n}}-\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}\right)^2,\end{aligned}$$

where the second line uses the fact that $\{z_{\ell,i}\}_\ell$ are independent standard normal random variables. By adding and subtracting terms, we have

$$\begin{aligned}\mathbb{E}\left(\frac{\|\phi(h^{n,L}_{u-1})\|}{\sqrt{n}}-\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}\right)^2 \leq{}& 2\mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{n,L}_{u-1,j})}-\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})}\right)^2\\ &+2\mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})}-\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}\right)^2. \quad (11)\end{aligned}$$

For the first term on the right hand side, using Minkowski's inequality, we have

$$\begin{aligned}&\mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{n,L}_{u-1,j})}-\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})}\right)^2\\ &\leq \mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\phi(h^{n,L}_{u-1,j})-\phi(h^{L}_{u-1,j})\right)^2}\right)^2\\ &\leq \frac{K_1^2}{n}\sum_{j=1}^{n}\mathbb{E}(h^{n,L}_{u-1,j}-h^{L}_{u-1,j})^2 = K_1^2\mathbb{E}(h^{n,L}_{u-1,i}-h^{L}_{u-1,i})^2.\end{aligned}$$

For the second term on the right hand side of equation 11, we have

$$\begin{aligned}&\mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})}-\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}\right)^2\\ &= \mathbb{E}\left(\frac{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\phi^2(h^{L}_{u-1,j})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi^2(h^{L}_{u-1,j})}+\sqrt{\mathbb{E}\phi^2(h^{L}_{u-1,i})}}\right)^2\\ &\leq C\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\phi^2(h^{L}_{u-1,j})-\mathbb{E}\phi^2(h^{L}_{u-1,j})\right]^2 \leq C/n,\end{aligned}$$

where the last line uses the independence of $\{h^{L}_{u-1,j}\}_j$ and Lemma 3. Therefore, we obtain

$$\mathbb{E}(h^{n,L}_{\ell,i}-h^{L}_{\ell,i})^2 \leq \frac{C_{K_1}}{L}\sum_{u=0}^{\ell-1}\mathbb{E}(h^{n,L}_{u,i}-h^{L}_{u,i})^2+\frac{C}{n}.$$

By discrete Gronwall's inequality (Lemma 2), we have the desired result. □

Next, to analyze the limit of $h^{L}_{\ell,i}$ as $L\to\infty$, we omit the subscript i and view

$$\sqrt{\frac{T}{L}}z_\ell = w\left(\frac{\ell T}{L}\right)-w\left(\frac{(\ell-1)T}{L}\right)$$

for a standard Brownian motion w. Then we can write $h_\ell^L = h_{\ell T/L}^{(L)}$, where

$$dh_t^{(L)} = \sqrt{\mathbb{E}\phi^2(h_{t_L}^{(L)})}\, dw_t$$

and $t_L := \lfloor \frac{t}{T/L} \rfloor \frac{T}{L}$ for $t \in [0, T]$. Consider the McKean–Vlasov process

$$dh_t = \sqrt{\mathbb{E}\phi^2(h_t)}\, dw_t.$$

Then $\{h_t^{(L)}\}$ is just the Euler–Maruyama discretization for $\{h_t\}$ with step size $\Delta t = T/L$.

The following is a standard result (see e.g. [26, Section I.1]) and we only provide a sketch of the proof.

Proposition 6. *There exists a unique* $\{h_t\}$ *and*

$$\sup_{0 \le t \le T} \mathbb{E}h_t^2 < \infty, \quad \mathbb{E}[h_t - h_{t_L}]^2 \le C(t - t_L) \le C/L. \tag{12}$$

Proof of Proposition 6. The evolution of h_t can be written as

$$dh_t = \sigma(\mu_t)\, dw_t, \quad h_0 \sim \mathcal{N}(0, \|\boldsymbol{x}\|^2/d),$$

where $\mu_t = \text{Law}(h_t)$ and

$$\sigma(\nu) := \sqrt{\int \phi^2(x)\, \nu(dx)} \tag{13}$$

for $\nu \in \mathcal{P}(\mathbb{R})$. Note that for any $X \sim \mu \in \mathcal{P}(\mathbb{R})$ and $Y \sim \nu \in \mathcal{P}(\mathbb{R})$, using Minkowski's inequality we have

$$|\sigma(\mu) - \sigma(\nu)| = |\sqrt{\mathbb{E}\phi^2(X)} - \sqrt{\mathbb{E}\phi^2(Y)}| \le \sqrt{\mathbb{E}[\phi(X) - \phi(Y)]^2}.$$

By Lipschitz property of ϕ, we have

$$|\sigma(\mu) - \sigma(\nu)| \le C W_2(\mu, \nu), \tag{14}$$

where $W_2(\cdot, \cdot)$ is the Wasserstein metric on $\mathcal{P}(\mathbb{R})$. Therefore, σ is a Lipschitz function. Now let

$$\mathcal{M} := \{\mu \in \mathcal{P}(\mathbb{C}([0, T] : \mathbb{R})) : \sup_{0 \le t \le T} \int x^2\, \mu_t(dx) < \infty\}. \tag{15}$$

For $\mu \in \mathcal{M}$, consider the process

$$dX_t = \sigma(\mu_t)\, dw_t, \quad X_0 = h_0. \tag{16}$$

It is well-defined and Law $(X) \in \mathcal{M}$, by Lipschitz property of ϕ. Denote the map from $\mu \in \mathcal{M}$ to Law$(X) \in \mathcal{M}$ by Γ. For $\mu, \nu \in \mathcal{M}$, denote the Wasserstein metric by

$$W_{2,t}(\mu, \nu) := \inf\{\left(\mathbb{E}[\sup_{u \le t} |X_u - Y_u|^2]\right)^{1/2} : \text{Law}(X) = \mu, \text{Law}(Y) = \nu\}. \tag{17}$$

Now given $\mu, \nu \in \mathcal{M}$, let

$$dX_t = \sigma(\mu_t)\, dw_t, \quad dY_t = \sigma(\nu_t)\, dw_t, \quad X_0 = Y_0 = h_0. \tag{18}$$

Then using Doob's maximal inequality, we have

$$\begin{aligned} W_{2,t}^2(\Gamma(\mu), \Gamma(\nu)) &\le \mathbb{E}[\sup_{u \le t} |X_u - Y_u|^2] = \mathbb{E}[\sup_{u \le t} | \int_0^u [\sigma(\mu_s) - \sigma(\nu_s)]\, dw_s|^2] \\ &\le 4\mathbb{E}| \int_0^t [\sigma(\mu_s) - \sigma(\nu_s)]\, dw_s|^2 = 4 \int_0^t [\sigma(\mu_s) - \sigma(\nu_s)]^2\, ds \\ &\le C \int_0^t W_2^2(\mu_s, \nu_s)\, ds \le C \int_0^t W_{2,s}^2(\mu, \nu)\, ds. \end{aligned}$$

Existence and uniqueness of $\{h_t\}$ then follows from standard arguments (cf. [26, Section I.1]). The first estimate in equation 12 follows from standard arguments on observing that ϕ is Lipscthiz and hence has linear growth. From this we immediately get the second estimate in equation 12. □

The following result quantifies the error as $L \to \infty$. This is not the stronger result one would usually get for Euler–Maruyama approximations. But it is sufficient for our use and also will be used in later inductive arguments for traning steps.

Proposition 7. *For all $L \geq 1$,*

$$\sup_{\ell=0,1,\dots,L} \mathbb{E}[h_\ell^L - h_{\ell T/L}]^2 = \sup_{\ell=0,1,\dots,L} \mathbb{E}[h_{\ell T/L}^{(L)} - h_{\ell T/L}]^2 \leq C/L. \tag{19}$$

Proof of Proposition 7. Let $s_L := \lfloor \frac{s}{T/L} \rfloor \frac{T}{L}$. Since $h_0^{(L)} = h_0$, we have

$$\begin{aligned}
&\mathbb{E}[h_\ell^L - h_{\ell T/L}]^2 = \mathbb{E}[h_{\ell T/L}^{(L)} - h_{\ell T/L}]^2 \\
&= \int_0^{\ell T/L} |\sqrt{\mathbb{E}\phi^2(h_{s_L}^{(L)})} - \sqrt{\mathbb{E}\phi^2(h_s)}|^2 \, ds \\
&\leq 2 \int_0^{\ell T/L} |\sqrt{\mathbb{E}\phi^2(h_{s_L}^{(L)})} - \sqrt{\mathbb{E}\phi^2(h_{s_L})}|^2 \, ds + 2 \int_0^{\ell T/L} |\sqrt{\mathbb{E}\phi^2(h_{s_L})} - \sqrt{\mathbb{E}\phi^2(h_s)}|^2 \, ds \\
&\leq \frac{C}{L} \sum_{u=0}^{\ell-1} \mathbb{E}[h_u^L - h_{uT/L}]^2 + \frac{C}{L},
\end{aligned}$$

where the last line uses the Lipschitz property in equation 14 and Lemma 6. It then follows from discrete Gronwall's inequality (Lemma 2) that

$$\mathbb{E}[h_\ell^L - h_{\ell T/L}]^2 \leq \frac{C}{L} e^{C\ell/L}. \tag{20}$$

This completes the proof. □

Combining Propositions 5 and 7, we get Proposition 2.

C.2 First Backward — Proof of Proposition 3

Suppose ϕ satisfies the assumption in Proposition 3 throughout.

Recall $\boldsymbol{g}_\ell$ in equation 4:

$$\boldsymbol{g}_L = n \frac{\partial f}{\partial \boldsymbol{h}_L} = \boldsymbol{v}, \quad \boldsymbol{g}_{\ell-1} = n \frac{\partial f}{\partial \boldsymbol{h}_\ell} = \boldsymbol{g}_\ell + \sqrt{\frac{T}{Ln}} \phi'(\boldsymbol{h}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell, \quad \forall \ell \in [L].$$

Under the gradient independence assumption, this is equivalent to

$$\begin{aligned}
g_{L,i}^{n,L} &= v_i, \\
g_{\ell-1,i}^{n,L} &= g_{\ell,i}^{n,L} + \sqrt{\frac{T}{Ln}} \phi'(h_{\ell-1,i}^{n,L}) \|\boldsymbol{g}_\ell^{n,L}\| \tilde{z}_{\ell,i},
\end{aligned}$$

where $\{\tilde{z}_{\ell,i}\}$ are independent standard normal random variables and also independent of $\{z_{\ell,i}\}$.

We note that the evolution of the first backward $g_{\ell-1,i}^{n,L}$ is very similar to that of the first forward $h_{\ell,i}^{n,L}$, except that the last term involves an extra term $\phi'(h_{\ell-1,i}^{n,L})$ and it depends on $\|\boldsymbol{g}_\ell^{n,L}\|$ without through the activation function ϕ. Therefore, the proof of Proposition 3 is very similar to that of Proposition 2, thanks to the assumption that ϕ' is Lipschitz. Hence we will only state the following results and omit most proofs.

First we want to show as $n \to \infty$, $\{g_{\ell,i}^{n,L}\}$ converges to

$$g_{\ell-1,i}^L = g_{\ell,i}^L + \sqrt{\frac{T}{L}} \phi'(h_{\ell-1,i}^L) \sqrt{\mathbb{E}(g_{\ell,i}^L)^2} \tilde{z}_{\ell,i}.$$

Lemma 4. *For each $i \in \mathbb{N}$, $\sup_{L \geq 1} \sup_{\ell=0,\dots,L} \mathbb{E}[g_{\ell,i}^L]^4 < \infty$ and $\inf_{L \geq 1} \inf_{\ell=0,\dots,L} \mathbb{E}(g_{\ell,i}^L)^2 > 0$.*

Proof of Lemma 4. Proof of the first assertion is omitted. For the second assertion, since $g_{L,i}^{n,L} = v_i$ is independent of $\{h_{\ell,i}^L\}$ and $\{\tilde z_{\ell,i}\}$, we have

$$\mathbb{E}(g_{\ell,i}^L)^2 \geq \mathrm{Var}(g_{\ell,i}^L) \geq \mathrm{Var}(v_i) = C > 0.$$

This completes the proof. □

Proposition 8. *For each $i \in \mathbb{N}$,*

$$\sup_{L\geq 1}\sup_{\ell=0,\dots,L} \mathbb{E}(g_{\ell,i}^{n,L} - g_{\ell,i}^L)^2 \leq C/n.$$

Next, to analyze the limit of $g_{\ell,i}^L$ as $L \to \infty$, we omit the subscript i and view

$$\sqrt{\frac{T}{L}}\tilde z_\ell = \tilde w\left(\frac{(\ell-1)T}{L}\right) - \tilde w\left(\frac{\ell T}{L}\right)$$

for a standard Brownian motion $\tilde w$ that goes backward in time. Then we can write $g_\ell^L = g_{\ell T/L}^{(L)}$, where

$$dg_t^{(L)} = \phi'(h_{\tilde t_L}^{(L)})\sqrt{\mathbb{E}(g_{\tilde t_L}^{(L)})^2}\, d\tilde w_t$$

and $\tilde t_L := \lceil \frac{t}{T/L} \rceil \frac{T}{L}$. Consider the McKean–Vlasov process (that goes backward in time)

$$dg_t = \sqrt{\mathbb{E}g_t^2}\, d\tilde w_t.$$

Then $\{g_t^{(L)}\}$ is just the Euler–Maruyama discretization for $\{g_t\}$ with step size $\Delta t = T/L$.

Proposition 9. *There exists a unique $\{g_t\}$ and*

$$\sup_{0\leq t\leq T} \mathbb{E}g_t^2 < \infty, \quad \mathbb{E}[g_t - g_{\tilde t_L}]^2 \leq C(\tilde t_L - t) \leq C/L. \tag{21}$$

Proposition 10. *For all $L \geq 1$,*

$$\sup_{\ell=0,1,\dots,L} \mathbb{E}[g_\ell^L - g_{\ell T/L}]^2 = \sup_{\ell=0,1,\dots,L} \mathbb{E}[g_{\ell T/L}^{(L)} - g_{\ell T/L}]^2 \leq C/L. \tag{22}$$

Combining Propositions 8 and 10, we get Proposition 3.

D Gradient Computation

In this section we derive the forward propagation and backward propagation processes after gradient updates. Recall that we have defined the ResNet as follows:

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

Then we consider the backward propagation of gradients as follows:

$$\boldsymbol{g}_L = \frac{\sqrt{n}}{\alpha}\frac{\partial f}{\partial \boldsymbol{h}_L} = \boldsymbol{v}, \qquad \boldsymbol{g}_{\ell-1} = \frac{\sqrt{n}}{\alpha}\frac{\partial f}{\partial \boldsymbol{h}_{\ell-1}} = \boldsymbol{g}_\ell + \sqrt{\frac{T}{Ln}}\phi'(\boldsymbol{h}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell, \quad \forall \ell \in [L].$$

The gradients of f w.r.t. trainable parameters are given by

$$\frac{\partial f}{\partial \boldsymbol{v}} = \frac{\alpha}{\sqrt{n}}\boldsymbol{h}_L, \qquad \frac{\partial f}{\partial \boldsymbol{W}_\ell} = \sqrt{\frac{T}{Ln}}\frac{\partial f}{\partial \boldsymbol{h}_\ell}\phi(\boldsymbol{h}_{\ell-1})^\top, \qquad \frac{\partial f}{\partial \boldsymbol{U}} = \frac{1}{\sqrt{d}}\frac{\partial f}{\partial \boldsymbol{h}_0}\boldsymbol{x}^\top.$$

Given a loss function $\mathcal{L}$, the SGD with a single sample is given by

$$\begin{aligned}
\boldsymbol{v}^+ &= \boldsymbol{v} - \eta\mathcal{L}'(f,y)\frac{\alpha}{\sqrt{n}}\boldsymbol{h}_L \\
\boldsymbol{W}_\ell^+ &= \boldsymbol{W}_\ell - \eta\mathcal{L}'(f,y)\sqrt{\frac{T}{Ln}}\frac{\partial f}{\partial \boldsymbol{h}_\ell}\phi(\boldsymbol{h}_{\ell-1})^\top = \boldsymbol{W}_\ell - \eta\frac{\alpha}{\sqrt{n}}\mathcal{L}'(f,y)\sqrt{\frac{T}{Ln}}\boldsymbol{g}_\ell\phi(\boldsymbol{h}_{\ell-1})^\top \\
\boldsymbol{U}^+ &= \boldsymbol{U} - \eta\mathcal{L}'(f,y)\frac{1}{\sqrt{d}}\frac{\partial f}{\partial \boldsymbol{h}_0}\boldsymbol{x}^\top = \boldsymbol{U} - \eta\frac{\alpha}{\sqrt{n}}\mathcal{L}'(f,y)\frac{1}{\sqrt{d}}\boldsymbol{g}_0\boldsymbol{x}^\top.
\end{aligned}$$

Then, after k step gradient updates, the forward propagation becomes:

$$\begin{aligned}
\boldsymbol{h}_0^{(k)} &= \frac{1}{\sqrt{d}}\boldsymbol{U}^{(k)}\boldsymbol{x}^{(k)} \\
&= \frac{1}{\sqrt{d}}\left[\boldsymbol{U} - \eta\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{1}{\sqrt{d}}\frac{\partial f^{(i)}}{\partial \boldsymbol{h}_0^{(i)}}\boldsymbol{x}^{(i)\top}\right]\boldsymbol{x}^{(k)} \\
&= \frac{1}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x}^{(k)} - \eta\frac{\alpha}{\sqrt{n}}\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(k)}\right\rangle}{d}\boldsymbol{g}_0^{(i)} \\
\boldsymbol{h}_\ell^{(k)} &= \boldsymbol{h}_{\ell-1}^{(k)} + \sqrt{\frac{T}{Ln}}\boldsymbol{W}_\ell^{(k)}\phi(\boldsymbol{h}_{\ell-1}^{(k)}) \\
&= \boldsymbol{h}_{\ell-1}^{(k)} + \sqrt{\frac{T}{Ln}}\left[\boldsymbol{W}_\ell - \eta\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\sqrt{\frac{T}{Ln}}\frac{\partial f^{(i)}}{\partial \boldsymbol{h}_\ell^{(i)}}\phi(\boldsymbol{h}_{\ell-1}^{(i)})^\top\right]\phi(\boldsymbol{h}_{\ell-1}^{(k)}) \\
&= \boldsymbol{h}_{\ell-1}^{(k)} - \eta\frac{\alpha}{\sqrt{n}}\frac{T}{L}\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\left\langle \phi(\boldsymbol{h}_{\ell-1}^{(i)}), \phi(\boldsymbol{h}_{\ell-1}^{(k)})\right\rangle}{n}\boldsymbol{g}_\ell^{(i)} + \sqrt{\frac{T}{Ln}}\boldsymbol{W}_\ell\phi(\boldsymbol{h}_{\ell-1}^{(k)}) \\
f^{(k)} &= \frac{\alpha}{\sqrt{n}}\left\langle \boldsymbol{v}^{(k)}, \boldsymbol{h}_L^{(k)}\right\rangle \\
&= \frac{\alpha}{\sqrt{n}}\left[\boldsymbol{v} - \eta\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\alpha}{\sqrt{n}}\boldsymbol{h}_L^{(i)}\right]^\top\boldsymbol{h}_L^{(k)} \\
&= \frac{\alpha}{\sqrt{n}}\boldsymbol{v}^\top\boldsymbol{h}_L^{(k)} - \eta\alpha^2\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\left\langle \boldsymbol{h}_L^{(i)}, \boldsymbol{h}_L^{(k)}\right\rangle}{n}
\end{aligned}$$

Consequentially, the backward propagation becomes

$$\begin{aligned}
\boldsymbol{g}_L^{(k)} &= \boldsymbol{v}^{(k)} = \boldsymbol{v} - \eta\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\alpha}{\sqrt{n}}\boldsymbol{h}_L^{(i)} \\
\boldsymbol{g}_{\ell-1}^{(k)} &= \boldsymbol{g}_\ell^{(k)} + \sqrt{\frac{T}{Ln}}\phi'(\boldsymbol{h}_{\ell-1}^{(k)}) \odot \boldsymbol{W}_\ell^{(k)\top}\boldsymbol{g}_\ell^{(k)} \\
&= \boldsymbol{g}_\ell^{(k)} + \sqrt{\frac{T}{Ln}}\phi'(\boldsymbol{h}_{\ell-1}^{(k)}) \odot \left[\boldsymbol{W}_\ell - \eta\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\sqrt{\frac{T}{Ln}}\frac{\partial f^{(i)}}{\partial \boldsymbol{h}_\ell^{(i)}}\phi(\boldsymbol{h}_{\ell-1}^{(i)})^\top\right]^\top\boldsymbol{g}_\ell^{(k)} \\
&= \boldsymbol{g}_\ell^{(k)} - \eta\frac{\alpha}{\sqrt{n}}\frac{T}{L}\sum_{i=0}^{k-1}\mathcal{L}'(f^{(i)}, y^{(i)})\frac{\left\langle \boldsymbol{g}_\ell^{(i)}, \boldsymbol{g}_\ell^{(k)}\right\rangle}{n}\left[\phi(\boldsymbol{h}_{\ell-1}^{(i)}) \odot \phi'(\boldsymbol{h}_{\ell-1}^{(k)})\right] \\
&\quad + \sqrt{\frac{T}{Ln}}\phi'(\boldsymbol{h}_{\ell-1}^{(k)}) \odot \boldsymbol{W}_\ell^\top\boldsymbol{g}_\ell^{(k)}
\end{aligned}$$

E Depth convergence — Proof of Theorem 1

In this section we start from Proposition 4, neglecting the last higher order term in equation 6, to prove the convergence in Theorem 1 and show that the rate of convergence as $L \to \infty$ is $1/L$. Recall that we assume $\mathcal{L}'$, ϕ, and ϕ' are Lipschitz continuous, and

$$\{\boldsymbol{\Sigma}_{t,k\times k}, \Theta_{t,k\times k}\}_{t\in[0,T]} \text{ are uniformly strictly positive definite for each } k \in \mathbb{N} \tag{23}$$

for the feature learning dynamics system. For ease of presentation, we consider the one sample case.

The limit as $n \to \infty$ for the K-th iteration can be written as

$$\begin{aligned}
h_\ell^{(K),L} &= h_{\ell-1}^{(K),L} - \eta_0 \frac{T}{L} \sum_{k=0}^{K-1} \mathcal{L}'(k,L) g_\ell^{(k),L} \mathbb{E}(\phi(h_{\ell-1}^{(k),L})\phi(h_{\ell-1}^{(K),L})) + \sqrt{\frac{T}{L}} z_\ell^{(K),L}, \\
g_{\ell-1}^{(K),L} &= g_\ell^{(K),L} - \eta_0 \frac{T}{L} \phi'(h_{\ell-1}^{(K),L}) \sum_{k=0}^{K-1} \mathcal{L}'(k,L) \phi(h_{\ell-1}^{(k),L}) \mathbb{E}(g_\ell^{(k),L}, g_\ell^{(K),L}) + \sqrt{\frac{T}{L}} \phi'(h_{\ell-1}^{(K),L}) \tilde{z}_\ell^{(K),L},
\end{aligned}$$

where $\{(z_\ell^{(k),L})_k, (\tilde{z}_\ell^{(k),L})_k : \ell = 1, \dots, L\}$ are independent Gaussian random vectors with mean 0 and variance-covariance matrix

$$\begin{aligned}
\mathrm{Cov}(z_\ell^{(k),L}, z_\ell^{(k'),L}) &= \mathbb{E}[\phi(h_{\ell-1}^{(k),L})\phi(h_{\ell-1}^{(k'),L})], \\
\mathrm{Cov}(\tilde{z}_\ell^{(k),L}, \tilde{z}_\ell^{(k'),L}) &= \mathbb{E}[g_\ell^{(k),L} g_\ell^{(k'),L}],
\end{aligned}$$

and

$$\begin{aligned}
h_0^{(K),L} &= \frac{\|\boldsymbol{x}\|}{\sqrt{d}} u(0) - \eta_0 \sum_{k=0}^{K-1} \mathcal{L}'(k,L) g_0^{(k),L} \frac{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}{d}, \\
g_L^{(K),L} &= v(0) - \eta_0 \sum_{k=0}^{K-1} \mathcal{L}'(k,L) h_L^{(k),L}, \\
\mathcal{L}'(k,L) &= \mathcal{L}'(\mathbb{E}[g_L^{(k),L} h_L^{(k),L}], y),
\end{aligned}$$

and $u(0), v(0)$ are standard Gaussians.

Letting $L \to \infty$, we expect to have

$$\begin{aligned}
dh_t^{(K)} &= -\eta_0 \sum_{k=0}^{K-1} \mathcal{L}'(k) g_t^{(k)} \mathbb{E}[\phi(h_t^{(k)})\phi(h_t^{(K)})] dt + dw_t^{(K)}, \quad \forall t \in [0,T], \\
dg_t^{(K)} &= -\eta_0 \phi'(h_t^{(K)}) \sum_{k=0}^{K-1} \mathcal{L}'(k) \phi(h_t^{(k)}) \mathbb{E}[g_t^{(k)} g_t^{(K)}] dt + \phi'(h_t^{(K)}) d\tilde{w}_t^{(K)}, \quad \forall t \in [0,T].
\end{aligned}$$

where $\{(w_t^{(k)})_k, (\tilde{w}_t^{(k)})_k : \ell = 1, \dots, L\}$ are indpendent Brownian motions with mean 0 and cross-variations

$$\begin{aligned}
d\langle w^{(k)}, w^{(k')} \rangle_t &= \mathbb{E}[\phi(h_t^{(k)})\phi(h_t^{(k')})] \, dt, \\
d\langle \tilde{w}^{(k)}, \tilde{w}^{(k')} \rangle_t &= \mathbb{E}[g_t^{(k)} g_t^{(k')}] \, dt,
\end{aligned}$$

and

$$\begin{aligned}
h_0^{(K)} &= \frac{\|\boldsymbol{x}\|}{\sqrt{d}} u(0) - \eta_0 \sum_{k=0}^{K-1} \mathcal{L}'(k) g_0^{(k)} \frac{\|\boldsymbol{x}\|^2}{d}, \\
g_T^{(K)} &= v(0) - \eta_0 \sum_{k=0}^{K-1} \mathcal{L}'(k) h_T^{(k)}, \\
\mathcal{L}'(k) &= \mathcal{L}'(\mathbb{E}[g_T^{(k)} h_T^{(k)}], y).
\end{aligned}$$

Remark 1. *(a) The evolution of $g_t^{(K)}$ is written for ease of notation and is interpreted backward from $t = T$ to $t = 0$. This is not to be confused with the classic notion of backward stochastic differential equations. The precise meaning, instead, is that $(g_t, \tilde{w}_t) = (\widehat{g}_{T-t}, \widehat{w}_{T-t})$ and*

$$d\widehat{g}_t^{(K)} = -\eta_0 \phi'(h_{T-t}^{(K)}) \sum_{k=0}^{K-1} \mathcal{L}'(k) \phi(h_{T-t}^{(k)}) \mathbb{E}[\widehat{g}_t^{(k)} \widehat{g}_t^{(K)}] dt + \phi'(h_{T-t}^{(K)}) d\widehat{w}_t^{(K)}, \quad \forall t \in [0,T].$$

(b) The first forward $\{h_t^{(0)}\}$ is adapted to the driven Brownian motions. Due to the backpropagation, $g_t^{(0)}$ and $\{h_t^{(k)}, g_t^{(k)}, k = 1, 2, \dots\}$ are not adapted any more. However, thanks to the deterministic diffusion coefficients in front of dw_t and $d\tilde{w}_t$, which are automatically adapted, the SDEs are well-posed, as is justified in Propositions 6 and 9 above and Proposition 11 below.

For ease of analysis, we write the above dynamics of $h_t := (h_t^{(0)}, \dots, h_t^{(K)})$ and $g_t := (g_t^{(0)}, \dots, g_t^{(K)})$ in the following more standard manner of McKean–Vlasov equations:

$$\begin{aligned} dh_t &= b_t\,dt + \sigma_t\,dW_t, \\ dg_t &= c_t\,dt + D_t\theta_t\,dB_t, \end{aligned}$$

where $b_t = (b_{t,k})_{k=0}^K$ and $c_t = (c_{t,k})_{k=0}^K$ are vectors given by

$$b_{t,k} = -\eta_0 \sum_{i=0}^{k-1} \mathcal{L}'(i) g_t^{(i)} \mathbb{E}[\phi(h_t^{(i)})\phi(h_t^{(k)})],$$

$$c_{t,k} := -\eta_0 \phi'(h_t^{(k)}) \sum_{i=0}^{k-1} \mathcal{L}'(i)\phi(h_t^{(i)})\mathbb{E}[g_t^{(i)} g_t^{(k)}],$$

D_t is a diagonal matrix given by

$$D_t = \text{diag}\{\phi'(h_t^{(0)}), \dots, \phi'(h_t^{(K)})\},$$

σ_t and θ_t are (the Cholesky decomposition) such that

$$\begin{aligned} \sigma_t\sigma_t^\top &= \Sigma_t := \mathbb{E}[(\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}))^\top (\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}))], \\ \theta_t\theta_t^\top &= \Theta_t := \mathbb{E}[(g_t^{(0)}, \dots, g_t^{(K)})^\top (g_t^{(0)}, \dots, g_t^{(K)})], \end{aligned}$$

and W_t and B_t are independent $(K+1)$-dimensional standard Brownian motions.

We note that the above system is nested: when K increases by 1, one simply adds one additional dimension to the evolution of h_t and g_t. This allows us to apply induction arguments in the proofs of later results. We also note that the existence of solutions to the above system is already guaranteed via the convergence of $n \to \infty$.

Denote by $\|\cdot\|$ the Frobenius norm of matrices (and vectors). Denote by $\lambda_k(A)$ the eigenvalues of a symmetric matrix A.

Lemma 5. *If $\{h_t^{(k)}, g_t^{(k)}, k = 0, 1, \dots, K\}$ is a solution, then*

$$\sup_{0\le t\le T} \mathbb{E}\|h_t\|^2 < \infty, \qquad \sup_{0\le t\le T} \mathbb{E}\|g_t\|^2 < \infty.$$

Proof of Lemma 5. We will prove by induction. The statement holds for $K = 0$ by Propositions 6 and 9.

Now suppose the statement holds for K, namely

$$\sup_{0\le t\le T} \sum_{k=0}^{K} \mathbb{E}[h_t^{(k)}]^2 < \infty, \qquad \sup_{0\le t\le T} \sum_{k=0}^{K} \mathbb{E}[g_t^{(k)}]^2 < \infty.$$

We will show that

$$\sup_{0\le t\le T} \mathbb{E}[h_t^{(K+1)}]^2 < \infty, \qquad \sup_{0\le t\le T} \mathbb{E}[g_t^{(K+1)}]^2 < \infty.$$

For $h_t^{(K+1)}$, using Cauchy-Schwarz inequality, Lipschitz property of ϕ, and induction assumption, we have

$$\begin{aligned} \mathbb{E}[h_t^{(K+1)}]^2 &\le C\mathbb{E}[h_0^{(K+1)}]^2 + C\sum_{k=0}^{K} \mathbb{E}\left(\int_0^t g_s^{(k)} \mathbb{E}[\phi(h_s^{(k)})\phi(h_s^{(K+1)})]\,ds\right)^2 + C\mathbb{E}[w_t^{(K+1)}]^2 \\ &\le C + C\sum_{k=0}^{K} \int_0^t \mathbb{E}[g_s^{(k)}]^2 \mathbb{E}\phi^2(h_s^{(k)})\mathbb{E}\phi^2(h_s^{(K+1)})\,ds + C\int_0^t \mathbb{E}\phi^2(h_s^{(K+1)})\,ds \\ &\le C + C\int_0^t \mathbb{E}[h_s^{(K+1)}]^2\,ds. \end{aligned}$$

It then follows from Gronwall's lemma that $\sup_{0\le t\le T} \mathbb{E}[h_t^{(K+1)}]^2 < \infty$. Since ϕ' is bounded, using similar arguments as above we can get $\sup_{0\le t\le T} \mathbb{E}[g_t^{(K+1)}]^2 < \infty$. Therefore the statement holds for $K + 1$ and this completes the proof by induction. □

Proposition 11. *Pathwise uniqueness holds for* $\{h_t^{(k)}, g_t^{(k)}, k = 0, 1, \dots, K\}$.

Proof of Proposition 11. We will prove by induction. By Propositions 6 and 9, $h_t^{(0)}$ and $g_t^{(0)}$ are unique. So the statement holds for $K = 0$.

Now suppose the statement holds for K, namely $h_t^{(k)}$ and $g_t^{(k)}$, $k = 0, 1, \dots, K$, are unique. We will show that $h_t^{(k)}$ and $g_t^{(k)}$, $k = 0, 1, \dots, K+1$ are unique. Consider the solution $(h_t^{(k)}, g_t^{(k)})_{k=0}^{K+1}$ and any other solution $(\tilde{h}_t^{(k)}, \tilde{g}_t^{(k)})_{k=0}^{K+1}$. By the induction assumption on uniqueness, we must have $(h_t^{(k)}, g_t^{(k)})_{k=0}^{K} = (\tilde{h}_t^{(k)}, \tilde{g}_t^{(k)})_{k=0}^{K}$. Recall

$$\sigma_t \sigma_t^\top = \Sigma_t = \mathbb{E}[(\phi(h_t^{(0)}), \dots, \phi(h_t^{(K+1)}))^\top (\phi(h_t^{(0)}), \dots, \phi(h_t^{(K+1)}))]$$

and let

$$\begin{aligned}
\tilde{\sigma}_t \tilde{\sigma}_t^\top = \tilde{\Sigma}_t &= \mathbb{E}[(\phi(\tilde{h}_t^{(0)}), \dots, \phi(\tilde{h}_t^{(K+1)}))^\top (\phi(\tilde{h}_t^{(0)}), \dots, \phi(\tilde{h}_t^{(K+1)}))] \\
&= \mathbb{E}[(\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}), \phi(\tilde{h}_t^{(K+1)}))^\top (\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}), \phi(\tilde{h}_t^{(K+1)}))].
\end{aligned}$$

Write Σ_t in block matrix form

$$\begin{aligned}
\Sigma_{11,t} &= \mathbb{E}[(\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}))^\top (\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}))], \\
\Sigma_{21,t} = \Sigma_{12,t}^\top &= \mathbb{E}[\phi(h_t^{(K+1)})(\phi(h_t^{(0)}), \dots, \phi(h_t^{(K)}))], \\
\Sigma_{22,t} &= \mathbb{E}[\phi^2(h_t^{(K+1)})],
\end{aligned}$$

corresponding to coordinates $0, 1, \dots, K$ and $K+1$. Also write σ_t, $\tilde{\Sigma}_t$ and $\tilde{\sigma}_t$ is the similar way. Then by Cholesky decomposition, we have

$$\begin{aligned}
\sigma_{11,t}\sigma_{11,t}^\top &= \Sigma_{11,t}, & \tilde{\sigma}_{11,t} &= \sigma_{11,t}, \\
\sigma_{12,t} &= 0, & \tilde{\sigma}_{12,t} &= 0, \\
\sigma_{21,t}^\top &= \sigma_{11,t}^{-1}\Sigma_{12,t}, & \tilde{\sigma}_{21,t}^\top &= \tilde{\sigma}_{11,t}^{-1}\tilde{\Sigma}_{12,t} = \sigma_{11,t}^{-1}\tilde{\Sigma}_{12,t}, \\
\sigma_{22,t} &= \sqrt{\Sigma_{22,t} - \sigma_{21,t}\sigma_{21,t}^\top}, & \tilde{\sigma}_{22,t} &= \sqrt{\tilde{\Sigma}_{22,t} - \tilde{\sigma}_{21,t}\tilde{\sigma}_{21,t}^\top}.
\end{aligned}$$

By Lipschitz property of ϕ, we have

$$\|\Sigma_{12,t} - \tilde{\Sigma}_{12,t}\|^2 + \|\Sigma_{22,t} - \tilde{\Sigma}_{22,t}\|^2 \le C\mathbb{E}[h_t^{(K+1)} - \tilde{h}_t^{(K+1)}]^2. \tag{24}$$

By equation 23, there exits some $\varepsilon > 0$ such that all eigenvalues of Σ_t are at least ε. It then follows from the eigenvalue interlacing theorem (of principal submatrix) that all eigenvalues of $\Sigma_{11,t}$ are at least ε. Then we have

$$\|\sigma_{11,t}^{-1}\|^2 = \text{trace}((\sigma_{11,t}^{-1})^\top \sigma_{11,t}^{-1}) = \text{trace}(\Sigma_{11,t}^{-1}) = \sum_{k=0}^{K} \frac{1}{\lambda_k(\Sigma_{11,t})} \le \frac{K+1}{\varepsilon}. \tag{25}$$

Therefore

$$\|\sigma_{21,t} - \tilde{\sigma}_{21,t}\|^2 = \|\sigma_{11,t}^{-1}(\Sigma_{12,t} - \tilde{\Sigma}_{12,t})\|^2 \le \|\sigma_{11,t}^{-1}\|^2 \|\Sigma_{12,t} - \tilde{\Sigma}_{12,t}\|^2 \le C\mathbb{E}[h_t^{(K+1)} - \tilde{h}_t^{(K+1)}]^2. \tag{26}$$

where the last inequality uses equation 25 and equation 24. Similarly,

$$\begin{aligned}
&|\sigma_{21,t}\sigma_{21,t}^\top - \tilde{\sigma}_{21,t}\tilde{\sigma}_{21,t}^\top|^2 = |(\sigma_{21,t} - \tilde{\sigma}_{21,t})(\sigma_{21,t} + \tilde{\sigma}_{21,t})^\top|^2 \\
&= |(\sigma_{21,t} - \tilde{\sigma}_{21,t})\sigma_{11,t}^{-1}(\Sigma_{12,t} + \tilde{\Sigma}_{12,t})|^2 \le \|\sigma_{21,t} - \tilde{\sigma}_{21,t}\|^2 \|\sigma_{11,t}^{-1}\|^2 \|\Sigma_{12,t} + \tilde{\Sigma}_{12,t}\|^2 \\
&\le C\mathbb{E}[h_t^{(K+1)} - \tilde{h}_t^{(K+1)}]^2,
\end{aligned} \tag{27}$$

where we have used Lemma 5 to get $\|\Sigma_{12,t} + \tilde{\Sigma}_{12,t}\|^2 \le C$. Also, note that

$$\sigma_{22,t}^2 = \Sigma_{22,t} - \sigma_{21,t}\sigma_{21,t}^\top = \Sigma_{22,t} - \Sigma_{21,t}\Sigma_{11,t}^{-1}\Sigma_{12,t} \tag{28}$$

is the Schur complement of the block $\Sigma_{11,t}$ of the matrix Σ_t, so that its eigenvalues are at least ε as well. Therefore $\sigma_{22,t} \geq \sqrt{\varepsilon}$ and hence

$$\begin{aligned}\|\sigma_{22,t} - \tilde{\sigma}_{22,t}\|^2 = \left(\frac{\sigma_{22,t}^2 - \tilde{\sigma}_{22,t}^2}{\sigma_{22,t} + \tilde{\sigma}_{22,t}}\right)^2 &\leq \frac{1}{\varepsilon}[(\Sigma_{22,t} - \sigma_{21,t}\sigma_{21,t}^\top) - (\tilde{\Sigma}_{22,t} - \tilde{\sigma}_{21,t}\tilde{\sigma}_{21,t}^\top)]^2 \\ &\leq C|\Sigma_{22,t} - \tilde{\Sigma}_{22,t}|^2 + C|\sigma_{21,t}\sigma_{21,t}^\top - \tilde{\sigma}_{21,t}\tilde{\sigma}_{21,t}^\top|^2 \leq C\mathbb{E}[h_t^{(K+1)} - \tilde{h}_t^{(K+1)}]^2,\end{aligned} \tag{29}$$

where the last inequality uses equation 24 and equation 27. Now note that

$$\begin{aligned}h_u^{(K+1)} - \tilde{h}_u^{(K+1)} = -\eta_0 \sum_{k=0}^{K} \int_0^u \mathcal{L}'(k) g_s^{(k)} \mathbb{E}[\phi(h_s^{(k)})(\phi(h_s^{(K+1)}) - \phi(\tilde{h}_s^{(K+1)}))]\, ds \\ + \int_0^u (\sigma_{21,s} - \tilde{\sigma}_{21,s}, \sigma_{22,s} - \tilde{\sigma}_{22,s})\, dW_s.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[\sup_{u\leq t} |h_u^{(K+1)} - \tilde{h}_u^{(K+1)}|^2] \leq C \sum_{k=0}^{K} \mathbb{E}\left[\sup_{u\leq t}\left|\int_0^u \mathcal{L}'(k) g_s^{(k)} \mathbb{E}[\phi(h_s^{(k)})(\phi(h_s^{(K+1)}) - \phi(\tilde{h}_s^{(K+1)}))]\, ds\right|^2\right] \\ + C\mathbb{E}\left[\sup_{u\leq t}\left|\int_0^u (\sigma_{21,s} - \tilde{\sigma}_{21,s}, \sigma_{22,s} - \tilde{\sigma}_{22,s})\, dW_s\right|^2\right].\end{aligned} \tag{30}$$

Here using Cauchy-Schwarz inequality, we can bound the first term on the right side by

$$\begin{aligned}&C \sum_{k=0}^{K} \mathbb{E} \int_0^t \left| g_s^{(k)} \mathbb{E}[\phi(h_s^{(k)})(\phi(h_s^{(K+1)}) - \phi(\tilde{h}_s^{(K+1)}))]\right|^2 ds \\ &\leq C \int_0^t \mathbb{E}[\phi(h_s^{(K+1)}) - \phi(\tilde{h}_s^{(K+1)})]^2\, ds \leq C \int_0^t \mathbb{E}[\sup_{u\leq s} |h_u^{(K+1)} - \tilde{h}_u^{(K+1)}|^2]\, ds.\end{aligned}$$

Using Doob's maximal inequality, we can bound the second term on the right side of equation 30 by

$$\begin{aligned}C\mathbb{E}\left|\int_0^t (\sigma_{21,s} - \tilde{\sigma}_{21,s}, \sigma_{22,s} - \tilde{\sigma}_{22,s})\, dW_s\right|^2 &= C \int_0^t [\|\sigma_{21,s} - \tilde{\sigma}_{21,s}\|^2 + \|\sigma_{22,s} - \tilde{\sigma}_{22,s}\|^2]\, ds \\ \leq C \int_0^t \mathbb{E}[h_s^{(K+1)} - \tilde{h}_s^{(K+1)}]^2\, ds &\leq C \int_0^t \mathbb{E}[\sup_{u\leq s} |h_s^{(K+1)} - \tilde{h}_s^{(K+1)}|^2]\, ds.\end{aligned}$$

where the first inequality uses equation 26 and equation 29. Combining above three estimates, we have

$$\mathbb{E}[\sup_{u\leq t} |h_u^{(K+1)} - \tilde{h}_u^{(K+1)}|^2] \leq C \int_0^t \mathbb{E}[\sup_{u\leq s} |h_s^{(K+1)} - \tilde{h}_s^{(K+1)}|^2]\, ds.$$

It then follows from Gronwall's inequality that

$$\mathbb{E}[\sup_{u\leq T} |h_u^{(K+1)} - \tilde{h}_u^{(K+1)}|^2] = 0. \tag{31}$$

This gives uniqueness of $h_t^{(K+1)}$. Since ϕ' is bounded, similar arguments as above give uniqueness of $g_t^{(K+1)}$. Therefore the statement holds for $K+1$ and this completes the proof by induction. □

Before proving the convergence rate as $L \to \infty$, we will need the following two preparation results. Recall $t_L := \lfloor \frac{t}{T/L} \rfloor \frac{T}{L}$ and $\tilde{t}_L := \lceil \frac{t}{T/L} \rceil \frac{T}{L}$ are the times corresponding to the discrete step.

Lemma 6. *If $\{h_t^{(k)}, g_t^{(k)}, k = 0, 1, \ldots, \kappa\}$ is a solution, then*

$$\mathbb{E}\|h_t - h_{t_L}\|^2 \leq C(t - t_L) \leq C/L, \quad \mathbb{E}\|g_t - g_{\tilde{t}_L}\|^2 \leq C(\tilde{t}_L - t) \leq C/L. \tag{32}$$

Proof of Lemma 6. Using Lemma 5 and Lipscthiz property of ϕ, we can deduce

$$\mathbb{E}\|b_t\|^2 \le C, \qquad \|\sigma_t\|^2 = \text{trace}(\sigma_t\sigma_t^\top) = \text{trace}(\Sigma_t) \le C,$$

and similarly $\mathbb{E}\|c_t\|^2 \le C$ and $\|\theta_t\|^2 \le C$. These give the desired result. □

Recall the infinite width limit $h_\ell^L := (h_t^{(0),L}, \ldots, h_t^{(K),L})$ and $g_\ell^L := (g_t^{(0),L}, \ldots, g_t^{(K),L})$.

Lemma 7. $\sup\limits_{L\ge 1}\sup\limits_{\ell=0,\ldots,L} \mathbb{E}\|h_\ell^L\|^2 < \infty$ *and* $\sup\limits_{L\ge 1}\sup\limits_{\ell=0,\ldots,L} \mathbb{E}\|g_\ell^L\|^2 < \infty$.

Proof of Lemma 7. We will prove by induction. By Lemmas 3 and 4, the statement holds for $K = 0$. Now suppose the statement holds for K, namely

$$\sup_{L\ge 1}\sup_{\ell=0,\ldots,L}\sum_{k=0}^{K}\mathbb{E}[h_\ell^{(k),L}]^2 < \infty, \quad \sup_{L\ge 1}\sup_{\ell=0,\ldots,L}\sum_{k=0}^{K}\mathbb{E}[g_\ell^{(k),L}]^2 < \infty. \tag{33}$$

We will show that

$$\sup_{L\ge 1}\sup_{\ell=0,\ldots,L}\mathbb{E}[h_\ell^{(K+1),L}]^2 < \infty, \quad \sup_{L\ge 1}\sup_{\ell=0,\ldots,L}\mathbb{E}[g_\ell^{(K+1),L}]^2 < \infty. \tag{34}$$

From the evolution of $h_\ell^{(K+1),L}$ and independence of $z_\cdot^{(K+1),L}$, we have

$$\begin{aligned}\mathbb{E}[h_\ell^{(K+1),L}]^2 \le 3\mathbb{E}[h_0^{(K+1),L}]^2 &+ 3\mathbb{E}\left[\sum_{u=1}^{\ell}\eta_0\frac{T}{L}\sum_{k=0}^{K}\mathcal{L}'(k,L)g_u^{(k),L}\mathbb{E}(\phi(h_{u-1}^{(k),L})\phi(h_{u-1}^{(K+1),L}))\right]^2\\ &+ 3\mathbb{E}\left[\sum_{u=1}^{\ell}\sqrt{\frac{T}{L}}z_u^{(K+1),L}\right]^2\\ &\le C + \frac{C\ell}{L^2}\sum_{u=1}^{\ell}\mathbb{E}\phi^2(h_{u-1}^{(K+1),L}) + \frac{C}{L}\sum_{u=1}^{\ell}\mathbb{E}\phi^2(h_{u-1}^{(K+1),L})\\ &\le C + \frac{C}{L}\sum_{u=0}^{\ell-1}\mathbb{E}[h_u^{(K+1),L}]^2.\end{aligned}$$

It then follows from discrete Gronwall's lemma again that

$$\mathbb{E}[h_\ell^{(K+1),L}]^2 \le Ce^{C\ell/L}. \tag{35}$$

Therefore $\sup\limits_{L\ge 1}\sup\limits_{\ell=0,\ldots,L}\mathbb{E}[h_\ell^{(K+1),L}]^2 < \infty$. Similar arguments give $\sup\limits_{L\ge 1}\sup\limits_{\ell=0,\ldots,L}\mathbb{E}[g_\ell^{(K+1),L}]^2 < \infty$ and hence the statement also holds for $K+1$. This completes the proof by induction. □

Now we couple h_ℓ^L and g_ℓ^L with h_t and g_t respectively, and state our result on the convergence rate of $1/L$ as $L \to \infty$. Denote by $L_t := \lfloor \frac{t}{T/L} \rfloor$, $L_s := \lfloor \frac{s}{T/L} \rfloor$, $\tilde{L}_t := \lceil \frac{t}{T/L} \rceil$, and $\tilde{L}_s := \lceil \frac{s}{T/L} \rceil$ for $s, t \in [0, T]$. We can write $h_\ell^L = h_{\ell T/L}^{(L)}$ and $g_\ell^L = g_{\ell T/L}^{(L)}$, where $h_t^{(L)}$ and $g_t^{(L)}$ are continuous interpolations using the same Brownian motions W_t and B_t:

$$\begin{aligned} dh_t^{(L)} &= b_{L_t}^{(L)}\,dt + \sigma_{L_t}^{(L)}\,dW_t,\\ dg_t^{(L)} &= c_{\tilde{L}_t}^{(L)}\,dt + D_{\tilde{L}_t}^{(L)}\theta_{\tilde{L}_t}^{(L)}\,dB_t.\end{aligned}$$

Here $b_\ell^{(L)} = (b_{\ell,k}^{(L)})_{k=0}^K$ and $c_\ell^{(L)} = (c_{\ell,k}^{(L)})_{k=0}^K$ are vectors given by

$$b_{\ell,k}^{(L)} = -\eta_0\frac{T}{L}\sum_{i=0}^{k-1}\mathcal{L}'(i,L)g_\ell^{(i),L}\mathbb{E}(\phi(h_{\ell-1}^{(i),L})\phi(h_{\ell-1}^{(k),L})),$$

$$c_{\ell,k}^{(L)} := -\eta_0\frac{T}{L}\phi'(h_{\ell-1}^{(k),L})\sum_{i=0}^{k-1}\mathcal{L}'(i,L)\phi(h_{\ell-1}^{(i),L})\mathbb{E}[g_\ell^{(i),L}g_\ell^{(k),L}],$$

$D_\ell^{(L)}$ is a diagonal matrix given by

$$D_\ell^{(L)} = \mathrm{diag}\{\phi'(h_{\ell-1}^{(0),L}), \dots, \phi'(h_{\ell-1}^{(K),L})\},$$

and $\sigma_\ell^{(L)}$ and $\theta_\ell^{(L)}$ are (the Cholesky decomposition) such that

$$\begin{aligned}\sigma_\ell^{(L)}(\sigma_\ell^{(L)})^\top &= \Sigma_\ell^{(L)} := \mathbb{E}[(\phi(h_{\ell-1}^{(0),L}), \dots, \phi(h_{\ell-1}^{(K),L}))^\top(\phi(h_{\ell-1}^{(0),L}), \dots, \phi(h_{\ell-1}^{(K),L}))],\\ \theta_\ell^{(L)}(\theta_\ell^{(L)})^\top &= \Theta_\ell^{(L)} := \mathbb{E}[(g_\ell^{(0),L}, \dots, g_\ell^{(K),L})^\top(g_\ell^{(0),L}, \dots, g_\ell^{(K),L})].\end{aligned}$$

The following proposition says that the L^2 error decays at a rate of $1/L$ for the coupled difference between h_ℓ^L (resp. g_ℓ^L), the finite depth process at discrete step ℓ, and $h_{\ell T/L}$ (resp. $g_{\ell T/L}$), the corresponding infinite-depth process at time $\ell T/L$.

Proposition 12. *For all $L \geq 1$,*

$$\sup_{\ell=0,1,\dots,L} \mathbb{E}\|h_\ell^L - h_{\ell T/L}\|^2 \leq C/L, \qquad \sup_{\ell=0,1,\dots,L} \mathbb{E}\|g_\ell^L - g_{\ell T/L}\|^2 \leq C/L. \tag{36}$$

Proof of Proposition 12. We first note that the Lipschitz estimates in equation 26 and equation 29 still hold when comparing σ_{s_L} and $\sigma_{L_s}^{(L)}$, thanks to equation 23. We will again prove by induction. By Propositions 7 and 10, the statement holds for $K = 0$.

Now suppose the statement holds for K, namely

$$\sup_{\ell=0,\dots,L} \sum_{k=0}^K \mathbb{E}[h_\ell^{(k),L} - h_{\ell T/L}^{(k)}]^2 \leq C/L, \qquad \sup_{\ell=0,\dots,L} \sum_{k=0}^K \mathbb{E}[g_\ell^{(k),L} - g_{\ell T/L}^{(k)}]^2 \leq C/L. \tag{37}$$

We will show that

$$\sup_{\ell=0,\dots,L} \mathbb{E}[h_\ell^{(K+1),L} - h_{\ell T/L}^{(K+1)}]^2 \leq C/L, \qquad \sup_{\ell=0,\dots,L} \mathbb{E}[g_\ell^{(K+1),L} - g_{\ell T/L}^{(K+1)}]^2 \leq C/L. \tag{38}$$

Note that

$$\begin{aligned}\mathbb{E}[h_\ell^{(K+1),L} - h_{\ell T/L}^{(K+1)}]^2 \leq 3\mathbb{E}[h_0^{(K+1),L} - h_0^{(K+1)}]^2 + 3\mathbb{E}\left[\int_0^{\ell T/L} (b_{L_s,K+1}^{(L)} - b_{s,K+1})\, ds\right]^2\\ + 3\mathbb{E}\left[\int_0^{\ell T/L} (\sigma_{21,L_s}^{(L)} - \sigma_{21,s}, \sigma_{22,L_s}^{(L)} - \sigma_{22,s})\, dW_s\right]^2.\end{aligned}$$

By induction assumption,

$$\mathbb{E}[h_0^{(K+1),L} - h_0^{(K+1)}]^2 \leq C/L. \tag{39}$$

By Lemmas 5, 6, and 7 we have

$$\begin{aligned}&\mathbb{E}\left[\int_0^{\ell T/L} (b_{L_s,K+1}^{(L)} - b_{s,K+1})\, ds\right]^2\\ &\leq C\int_0^{\ell T/L} \mathbb{E}|b_{L_s,K+1}^{(L)} - b_{s_L,K+1}|^2\, ds + C\int_0^{\ell T/L} \mathbb{E}|b_{s_L,K+1} - b_{s,K+1}|^2\, ds\\ &\leq \frac{C}{L}\sum_{u=0}^{\ell-1} \mathbb{E}[h_u^{(K+1),L} - h_{uT/L}^{(K+1)}]^2 + \frac{C}{L}.\end{aligned}$$

By Lemmas 5, 6, and 7, and Lipschitz property of σ's, we have

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{\ell T/L} (\sigma^{(L)}_{21,L_s} - \sigma_{21,s}, \sigma^{(L)}_{22,L_s} - \sigma_{22,s})\, dW_s\right]^2 \\
&= \int_0^{\ell T/L} [\|\sigma^{(L)}_{21,L_s} - \sigma_{21,s}\|^2 + \|\sigma^{(L)}_{22,L_s} - \sigma_{22,s}\|^2]\, ds \\
&\leq 2\int_0^{\ell T/L} [\|\sigma^{(L)}_{21,L_s} - \sigma_{21,s_L}\|^2 + \|\sigma^{(L)}_{22,L_s} - \sigma_{22,s_L}\|^2]\, ds \\
&\quad + 2\int_0^{\ell T/L} [\|\sigma_{21,s_L} - \sigma_{21,s}\|^2 + \|\sigma_{22,s_L} - \sigma_{22,s}\|^2]\, ds \\
&\leq \frac{C}{L}\sum_{u=0}^{\ell-1} \mathbb{E}[h_u^{(K+1),L} - h^{(K+1)}_{uT/L}]^2 + \frac{C}{L}.
\end{aligned}$$

Combining the above estimates gives

$$\mathbb{E}[h_\ell^{(K+1),L} - h^{(K+1)}_{\ell T/L}]^2 \leq \frac{C}{L}\sum_{u=0}^{\ell-1} \mathbb{E}[h_u^{(K+1),L} - h^{(K+1)}_{uT/L}]^2 + \frac{C}{L}.$$

It then follows from discrete Gronwall's lemma that

$$\mathbb{E}[h_\ell^{(K+1),L} - h^{(K+1)}_{\ell T/L}]^2 \leq \frac{C}{L} e^{C\ell/L}. \tag{40}$$

Therefore $\sup_{\ell=0,1,\ldots,L} \mathbb{E}[h_\ell^{(K+1),L} - h^{(K+1)}_{\ell T/L}]^2 \leq C/L$. Similar arguments give $\sup_{\ell=0,1,\ldots,L} \mathbb{E}[g_\ell^{(K+1),L} - g^{(K+1)}_{\ell T/L}]^2 \leq C/L$ and hence the statement holds for $K+1$. This completes the proof by induction. □

Using Proposition 12, we get Theorem 1.

F Feature Learning Dynamics in the Infinite-Width Limit: Intuition

In this section, we provide intuition for Proposition 4, aimed at readers who may not be familiar with the Tensor Program framework.

To analyze how the feature space evolves during training, we focus on the *feature learning regime* with scaling $\alpha = 1/\sqrt{n}$ and learning rate $\eta = \eta_c n$, where $\eta_c > 0$ is a fixed constant. Under this setting, the forward and backward recursions take the form:

$$\begin{aligned}
\boldsymbol{h}_0^{(k)} =& \frac{1}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x}^{(k)} - \eta_c \sum_{i=0}^{k-1} \mathcal{L}'(f^{(i)}, y^{(i)}) \frac{\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(k)} \right\rangle}{d} \boldsymbol{g}_0^{(i)}, \\
\boldsymbol{h}_\ell^{(k)} =& \boldsymbol{h}_{\ell-1}^{(k)} - \eta_c \frac{T}{L} \sum_{i=0}^{k-1} \mathcal{L}'(f^{(i)}, y^{(i)}) \frac{\left\langle \phi(\boldsymbol{h}_{\ell-1}^{(i)}), \phi(\boldsymbol{h}_{\ell-1}^{(k)}) \right\rangle}{n} \boldsymbol{g}_\ell^{(i)} + \sqrt{\frac{T}{Ln}} \boldsymbol{W}_\ell \phi(\boldsymbol{h}_{\ell-1}^{(k)}), \\
f^{(k)} =& \frac{1}{n}\boldsymbol{v}^\top \boldsymbol{h}_L^{(k)} - \eta_c \sum_{i=0}^{k-1} \mathcal{L}'(f^{(i)}, y^{(i)}) \frac{\left\langle \boldsymbol{h}_L^{(i)}, \boldsymbol{h}_L^{(k)} \right\rangle}{n},
\end{aligned}$$

and the backward recursion:

$$\begin{aligned}
\boldsymbol{g}_L^{(k)} =& \boldsymbol{v} - \eta_c \sum_{i=0}^{k-1} \mathcal{L}'(f^{(i)}, y^{(i)}) \boldsymbol{h}_L^{(i)}, \\
\boldsymbol{g}_{\ell-1}^{(k)} =& \boldsymbol{g}_\ell^{(k)} - \eta_c \frac{T}{L} \sum_{i=0}^{k-1} \mathcal{L}'(f^{(i)}, y^{(i)}) \frac{\left\langle \boldsymbol{g}_\ell^{(i)}, \boldsymbol{g}_\ell^{(k)} \right\rangle}{n} \left[\phi(\boldsymbol{h}_{\ell-1}^{(i)}) \odot \phi'(\boldsymbol{h}_{\ell-1}^{(k)})\right] \\
&+ \sqrt{\frac{T}{Ln}}\, \phi'(\boldsymbol{h}_{\ell-1}^{(k)}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell^{(k)}.
\end{aligned}$$

Observe that the training process indeed defines a valid Tensor Program: each vector $\{\boldsymbol{h}_\ell^{(k)}, \boldsymbol{g}_\ell^{(k)}\}$ is generated sequentially from previously constructed vectors or from the initial random parameters $\{\boldsymbol{v}, \boldsymbol{W}_\ell, \boldsymbol{U}\}$ via standard TP operations (MatMul, Nonlin, Moment). Consequently, we may directly apply the Master Theorem of [34, Theorem 7.4] to characterize their infinite-width mean-field limits.

In particular, for any finite collection of valid TP vectors $\{\boldsymbol{h}_s\}_{s=1}^M$ and any sufficiently regular test function $\psi : \mathbb{R}^M \to \mathbb{R}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\boldsymbol{h}_{1,i}, \boldsymbol{h}_{2,i}, \dots, \boldsymbol{h}_{M,i}) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\big[\psi(Z^{\boldsymbol{h}_1}, Z^{\boldsymbol{h}_2}, \dots, Z^{\boldsymbol{h}_M})\big], \tag{41}$$

where each $Z^{\boldsymbol{h}}$ denotes the *mean-field variable* associated with $\boldsymbol{h}$, i.e., the limiting distribution of a typical coordinate of $\boldsymbol{h}$ as width grows.

First Forward Pass At initialization, the ResNet can be written as

$$\begin{aligned}
\boldsymbol{h}_0 &= \frac{1}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x}, \\
\boldsymbol{h}_\ell &= \boldsymbol{h}_{\ell-1} + \sqrt{\tfrac{\tau}{n}}\boldsymbol{W}_\ell \phi(\boldsymbol{h}_{\ell-1}), \\
f(\boldsymbol{x}) &= \frac{1}{n}\boldsymbol{v}^\top \boldsymbol{h}_L,
\end{aligned}$$

where $\tau := T/L$. Since $\boldsymbol{U}_{ij} \sim \mathcal{N}(0,1)$ are i.i.d., each coordinate of $\boldsymbol{h}_0$ is i.i.d. with distribution $Z^{\boldsymbol{h}_0} \sim \mathcal{N}(0, \|\boldsymbol{x}\|^2/d)$. Applying the nonlinearity ϕ element-wise preserves independence across coordinates, so $\phi(\boldsymbol{h}_0)$ also has i.i.d. coordinates. Similarly, because $\boldsymbol{W}_{\ell,ij} \sim \mathcal{N}(0,1)$ are i.i.d., each coordinate of $\frac{1}{\sqrt{n}}\boldsymbol{W}_\ell\phi(\boldsymbol{h}_{\ell-1})$ is approximately Gaussian with mean zero and variance $\frac{1}{n}\|\phi(\boldsymbol{h}_{\ell-1})\|^2$. By the Master Theorem, as $n \to \infty$, these coordinates converge in distribution to a mean-field variable $Z^{\boldsymbol{W}_\ell \phi_{\ell-1}}$ with variance $\mathbb{E}[\phi^2(Z^{\boldsymbol{h}_{\ell-1}})]$, where, inductively, we use the fact that each coordinate of $\boldsymbol{h}_\ell$ converges to a mean-field variable $Z^{\boldsymbol{h}_\ell}$ in the infinite-width limit. Finally, since $\boldsymbol{v}_i \sim \mathcal{N}(0,1)$ are i.i.d., the network output satisfies

$$f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{h}_{L,i} \xrightarrow[n\to\infty]{a.s.} \mathring{f} := \mathbb{E}[Z^{\boldsymbol{v}} Z^{\boldsymbol{h}_L}],$$

by the law of large numbers.

Hence, in the infinite-width limit, the ResNet converges to the mean-field process

$$\begin{aligned}
Z^{\boldsymbol{h}_0} &= \mathcal{N}\big(0, \|\boldsymbol{x}\|^2/d\big), \\
Z^{\boldsymbol{h}_\ell} &= Z^{\boldsymbol{h}_{\ell-1}} + \sqrt{\tau}\, Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}, \quad \forall \ell \in [L], \\
\mathring{f} &= \mathbb{E}[Z^{\boldsymbol{v}} Z^{\boldsymbol{h}_L}],
\end{aligned}$$

where $\{Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}\}_{\ell=1}^L$ are centered jointly Gaussian random variables with covariance

$$\mathrm{Cov}\big(Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}, Z^{\boldsymbol{W}_k \boldsymbol{\phi}_{k-1}}\big) = \delta_{\ell,k}\, \mathbb{E}\big[\phi(Z^{\boldsymbol{h}_{\ell-1}})\phi(Z^{\boldsymbol{h}_{k-1}})\big], \quad \forall\, \ell, k \in [L],$$

and independent of $Z^{\boldsymbol{v}} \sim \mathcal{N}(0,1)$.

First Backward Pass In the feature learning regime, the first backward recursion for computing gradients is

$$\begin{aligned}
\boldsymbol{g}_L &= \boldsymbol{v}, \\
\boldsymbol{g}_{\ell-1} &= \boldsymbol{g}_\ell + \sqrt{\tfrac{\tau}{n}}\, \phi'(\boldsymbol{h}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell.
\end{aligned}$$

Since $\boldsymbol{v}$ has i.i.d. standard Gaussian coordinates independent of the forward activations $\{\boldsymbol{h}_\ell\}$, the recursion in the mean-field limit begins with $Z^{\boldsymbol{g}_L} = Z^{\boldsymbol{v}} \sim \mathcal{N}(0,1)$. Because the output head $\boldsymbol{v}$ is not reused elsewhere, the *gradient independence assumption (GIA)* [33] holds in the infinite-width limit. This permits replacing $\boldsymbol{W}_\ell^\top$ in the backward pass with an independent copy $\widetilde{\boldsymbol{W}}_\ell^\top$, so that $\widetilde{\boldsymbol{W}}_\ell^\top \boldsymbol{g}_\ell$

is independent of the forward variables $\boldsymbol{h}_\ell$. The coordinates of $\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{W}}_\ell^\top \boldsymbol{g}_\ell$ are then approximately i.i.d. Gaussian, and by the Master Theorem converge to a mean-field random variable $Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}$. Thus, in the infinite-width limit, the backward recursion is characterized by

$$\begin{aligned} Z^{\boldsymbol{g}_L} &= Z^{\boldsymbol{v}} \sim \mathcal{N}(0,1), \\ Z^{\boldsymbol{g}_{\ell-1}} &= Z^{\boldsymbol{g}_\ell} + \sqrt{\tau}\,\phi'(Z^{\boldsymbol{h}_{\ell-1}})\,Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}, \quad \forall \ell \in [L], \end{aligned}$$

where $\{Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}\}_{\ell=1}^L$ are centered jointly Gaussian random variables with covariance

$$\mathrm{Cov}\big(Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}, Z^{\boldsymbol{W}_k^\top \boldsymbol{g}_k}\big) = \delta_{\ell,k}\,\mathbb{E}[Z^{\boldsymbol{g}_\ell} Z^{\boldsymbol{g}_k}], \quad \forall\,\ell, k \in [L].$$

Second Forward Pass After one gradient update, the forward pass with a new input $\bar{\boldsymbol{x}}$ is given by

$$\begin{aligned} \bar{\boldsymbol{h}}_0 &= \tfrac{1}{\sqrt{d}}\,\boldsymbol{U}\bar{\boldsymbol{x}} - \eta_c \mathcal{L}'(f,y)\tfrac{\langle \boldsymbol{x}, \bar{\boldsymbol{x}}\rangle}{d}\,\boldsymbol{g}_0, \\ \bar{\boldsymbol{h}}_\ell &= \bar{\boldsymbol{h}}_{\ell-1} - \tau\eta_c \mathcal{L}'(f,y)\tfrac{\langle \phi(\boldsymbol{h}_{\ell-1}), \phi(\bar{\boldsymbol{h}}_{\ell-1})\rangle}{n}\,\boldsymbol{g}_\ell + \sqrt{\tfrac{\tau}{n}}\,\boldsymbol{W}_\ell \phi(\bar{\boldsymbol{h}}_{\ell-1}). \end{aligned}$$

Since $f \to \mathring{f}$ in the infinite-width limit, continuity of $\mathcal{L}'$ ensures $\mathcal{L}'(f,y) \to \mathcal{L}'(\mathring{f},y)$. Hence,

$$Z^{\bar{\boldsymbol{h}}_0} = Z^{\boldsymbol{U}\bar{\boldsymbol{x}}} - \eta_c \mathcal{L}'(\mathring{f},y)\tfrac{\langle \boldsymbol{x}, \bar{\boldsymbol{x}}\rangle}{d} Z^{\boldsymbol{g}_0},$$

where $Z^{\boldsymbol{U}\bar{\boldsymbol{x}}}$ is centered Gaussian correlated with $Z^{\boldsymbol{U}\boldsymbol{x}}$, with covariance $\mathrm{Cov}(Z^{\boldsymbol{U}\boldsymbol{x}}, Z^{\boldsymbol{U}\bar{\boldsymbol{x}}}) = \frac{1}{d}\boldsymbol{x}^\top \bar{\boldsymbol{x}}$.

For hidden states, the inner product $\frac{1}{n}\big\langle \phi(\boldsymbol{h}_{\ell-1}), \phi(\bar{\boldsymbol{h}}_{\ell-1})\big\rangle$ is a Moment operation in the TP, and by the Master Theorem converges to $\mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}})\phi(Z^{\bar{\boldsymbol{h}}_{\ell-1}})]$.

Next consider $\frac{1}{\sqrt{n}}\boldsymbol{W}_\ell \phi(\bar{\boldsymbol{h}}_{\ell-1})$. If we adopt the decoupled analysis (replacing $\boldsymbol{W}_\ell^\top$ in the backward pass with $\tilde{\boldsymbol{W}}_\ell^\top$), then by the CLT heuristic the coordinates converge to a Gaussian random variable $Z^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}}$, correlated with $Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}$ with

$$\mathrm{Cov}\Big(Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}, Z^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}}\Big) = \mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}})\phi(Z^{\bar{\boldsymbol{h}}_{\ell-1}})].$$

This CLT heuristic is valid only because we assume $\widetilde{\boldsymbol{W}}_\ell^\top$ is used in the backward pass; otherwise, $\bar{\boldsymbol{h}}_\ell$ and $\boldsymbol{W}_\ell^\top$ are strongly correlated through $\boldsymbol{g}_{\ell-1}$. Before exploring this coupling scenario, the second forward pass under the decoupling scenario is described as follows:

$$\begin{aligned} Z^{\bar{\boldsymbol{h}}_0} =& Z^{\boldsymbol{U}\bar{\boldsymbol{x}}} - \eta_c \mathcal{L}'(\mathring{f},y)\tfrac{\langle \boldsymbol{x}, \bar{\boldsymbol{x}}\rangle}{d} Z^{\boldsymbol{g}_0} \\ Z^{\bar{\boldsymbol{h}}_\ell} =& Z^{\bar{\boldsymbol{h}}_{\ell-1}} - \tau\eta_c \mathcal{L}'(\mathring{f},y)\mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}})\phi(Z^{\bar{\boldsymbol{h}}_{\ell-1}})] Z^{\boldsymbol{g}_\ell} + \sqrt{\tau} Z^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}}. \end{aligned}$$

Now, we focus on the normal gradient update to expose the effect of the reuse of $\boldsymbol{W}_\ell$ in the backward pass. For intuition, set $\phi = \mathrm{id}$. Expanding $\bar{\boldsymbol{h}}_{\ell-1}$ yields

$$\bar{\boldsymbol{h}}_{\ell-1} = \bar{\boldsymbol{h}}_{\ell-2} + \tau a_{\ell-2}\boldsymbol{g}_\ell + \tau a_{\ell-2}\sqrt{\tfrac{\tau}{n}}\,\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell + \sqrt{\tfrac{\tau}{n}}\,\boldsymbol{W}_{\ell-1}\bar{\boldsymbol{h}}_{\ell-2},$$

where

$$a_\ell := -\eta_c \mathcal{L}'(f,y)\frac{\big\langle \phi(\boldsymbol{h}_\ell), \phi(\bar{\boldsymbol{h}}_\ell)\big\rangle}{n} \xrightarrow{\text{a.s.}} \mathring{a}_\ell := -\eta_c \mathcal{L}'(\mathring{f},y)\mathbb{E}[\phi(Z^{\boldsymbol{h}_\ell})\phi(Z^{\bar{\boldsymbol{h}}_\ell})],$$

since a_ℓ is a valid TP scalar. Substituting into $\sqrt{\frac{\tau}{n}}\,\boldsymbol{W}_\ell \bar{\boldsymbol{h}}_{\ell-1}$ yields

$$\sqrt{\frac{\tau}{n}}\boldsymbol{W}_\ell \bar{\boldsymbol{h}}_{\ell-1} = \sqrt{\frac{\tau}{n}}\boldsymbol{W}_\ell\left(\bar{\boldsymbol{h}}_{\ell-2} + \tau a_{\ell-2}\boldsymbol{g}_\ell + \sqrt{\frac{\tau}{n}}\boldsymbol{W}_{\ell-1}\bar{\boldsymbol{h}}_{\ell-2}\right) + \tau^2 a_{\ell-2}\frac{1}{n}\boldsymbol{W}_\ell \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell.$$

The i-th coordinate of the $\frac{1}{n}\boldsymbol{W}_\ell \boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell$ can be decomposed into

$$\boldsymbol{g}_{\ell,i} \cdot \frac{1}{n}\sum_j \boldsymbol{W}_{\ell,ij}^2 + \frac{1}{n}\sum_j \sum_{k \neq i} \boldsymbol{W}_{\ell,ij}\boldsymbol{W}_{\ell,kj}\boldsymbol{g}_{\ell,k}.$$

By the law of large numbers, the first term converges to $Z^{\boldsymbol{g}_\ell}$, and the second converges (by CLT) to a Gaussian field $\hat{Z}^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}}$. Notably, under GIA, we replace $\boldsymbol{W}_{\ell,ij}^2$ with $\boldsymbol{W}_{\ell,ij}\widetilde{\boldsymbol{W}}_{\ell,ij}$. Hence, the first term vanishes as $n \to \infty$ because of the independence and zero mean.

Thus, for $\phi = \mathrm{id}$,

$$\left[\sqrt{\tfrac{\tau}{n}}\,\boldsymbol{W}_\ell \bar{\boldsymbol{h}}_{\ell-1}\right]_i \xrightarrow{\text{a.s.}} \sqrt{\tau}\,\hat{Z}^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}} + \tau^2 \mathring{a}_{\ell-2}\, Z^{\boldsymbol{g}_\ell}.$$

Here $\hat{Z}^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}}$ is Gaussian, correlated with $Z^{\boldsymbol{W}_\ell \boldsymbol{\phi}_{\ell-1}}$ from the first forward pass with covariance $\mathbb{E}[\phi(Z^{\boldsymbol{h}_{\ell-1}})\phi(Z^{\bar{\boldsymbol{h}}_{\ell-1}})]$.

Putting the pieces together, the mean-field feature dynamics after one gradient step are

$$\begin{aligned} Z^{\bar{\boldsymbol{h}}_0} &= Z^{\boldsymbol{U}\bar{\boldsymbol{x}}} - \eta_c \mathcal{L}'(\mathring{f}, y) \tfrac{\langle \boldsymbol{x}, \bar{\boldsymbol{x}} \rangle}{d}\, Z^{\boldsymbol{g}_0}, \\ Z^{\bar{\boldsymbol{h}}_\ell} &= Z^{\bar{\boldsymbol{h}}_{\ell-1}} - \tau \eta_c \mathcal{L}'(\mathring{f}, y) \mathbb{E}[Z^{\boldsymbol{h}_{\ell-1}} Z^{\bar{\boldsymbol{h}}_{\ell-1}}]\, Z^{\boldsymbol{g}_\ell} + \sqrt{\tau}\,\hat{Z}^{\boldsymbol{W}_\ell \bar{\boldsymbol{\phi}}_{\ell-1}} \\ &\quad - \tau^2 \eta_c \mathcal{L}'(\mathring{f}, y) \mathbb{E}[Z^{\boldsymbol{h}_{\ell-1}} Z^{\bar{\boldsymbol{h}}_{\ell-1}}]\, Z^{\boldsymbol{g}_\ell}. \end{aligned}$$

Comparing this result with the decoupling scenario, the final correction term reflects the additional interaction in both forward and backward paths due to the reuse of $\boldsymbol{W}_\ell$. Moreover, its scaling is τ^2 due to depth-adaptive ResNet normalization. By the Euler–Maruyama convergence perspective, this higher-order term vanishes as $L \to \infty$, provided the other quantities remain well behaved.

Second Backward Pass Analogously, we describe the backward propagation after one step of gradient descent in the mean-field limit. Given the second forward pass with input $\bar{\boldsymbol{x}}$, the second backward recursion is

$$\begin{aligned} \bar{\boldsymbol{g}}_L &= \boldsymbol{v} - \eta_c \mathcal{L}'(f, y)\, \boldsymbol{h}_L, \\ \bar{\boldsymbol{g}}_{\ell-1} &= \bar{\boldsymbol{g}}_\ell - \tau \eta_c \mathcal{L}'(f, y)\, \tfrac{\langle \boldsymbol{g}_\ell, \bar{\boldsymbol{g}}_\ell \rangle}{n} \left[\phi(\boldsymbol{h}_{\ell-1}) \odot \phi'(\bar{\boldsymbol{h}}_{\ell-1})\right] + \sqrt{\tfrac{\tau}{n}}\, \phi'(\bar{\boldsymbol{h}}_{\ell-1}) \odot \boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell. \end{aligned}$$

Assume we decouple the two backward passes by replacing $\boldsymbol{W}_\ell^\top$ with an independent copy $\widetilde{\boldsymbol{W}}_\ell^\top$ in the second backward recursion. Then, in the mean-field limit, the coordinates of $\frac{1}{\sqrt{n}} \widetilde{\boldsymbol{W}}_\ell^\top \bar{\boldsymbol{g}}_\ell$ converge (by the Master Theorem) to a centered Gaussian random variable $Z^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell}$, correlated with $Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}$ from the first backward pass via

$$\mathrm{Cov}\big(Z^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell}, Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}\big) \;=\; \mathbb{E}\big[Z^{\bar{\boldsymbol{g}}_\ell}\, Z^{\boldsymbol{g}_\ell}\big]\,.$$

Passing to the limit (and using continuity of $\mathcal{L}'$ so that $\mathcal{L}'(f, y) \to \mathcal{L}'(\mathring{f}, y)$), we obtain

$$\begin{aligned} Z^{\bar{\boldsymbol{g}}_L} &= Z^{\boldsymbol{g}_L} - \eta_c \mathcal{L}'(\mathring{f}, y)\, Z^{\boldsymbol{h}_L}, \\ Z^{\bar{\boldsymbol{g}}_{\ell-1}} &= Z^{\bar{\boldsymbol{g}}_\ell} - \tau \eta_c \mathcal{L}'(\mathring{f}, y)\, \mathbb{E}\big[Z^{\boldsymbol{g}_\ell} Z^{\bar{\boldsymbol{g}}_\ell}\big]\; \phi(Z^{\boldsymbol{h}_{\ell-1}})\, \phi'(Z^{\bar{\boldsymbol{h}}_{\ell-1}}) + \sqrt{\tau}\, \phi'(Z^{\bar{\boldsymbol{h}}_{\ell-1}})\, Z^{\tilde{\boldsymbol{W}}_\ell^\top \bar{\boldsymbol{g}}_\ell}. \end{aligned}$$

When the same weights $\boldsymbol{W}_\ell^\top$ are reused in both backward passes, additional correlations appear. For intuition, set $\phi = \mathrm{id}$ so $\phi' \equiv 1$. The second backward state expands as

$$\bar{\boldsymbol{g}}_\ell = \bar{\boldsymbol{g}}_{\ell+1} + \tau b_{\ell+1}\, \boldsymbol{h}_{\ell-1} + \tau b_{\ell+1} \sqrt{\tfrac{\tau}{n}}\, \boldsymbol{W}_\ell \boldsymbol{h}_{\ell-1} + \sqrt{\tfrac{\tau}{n}}\, \boldsymbol{W}_{\ell+1}^\top \bar{\boldsymbol{g}}_{\ell+1},$$

with

$$b_\ell \;:=\; -\,\eta_c\, \mathcal{L}'(f, y)\, \tfrac{\langle \boldsymbol{g}_\ell, \bar{\boldsymbol{g}}_\ell \rangle}{n} \;\xrightarrow{\text{a.s.}}\; \mathring{b}_\ell \;:=\; -\,\eta_c\, \mathcal{L}'(\mathring{f}, y)\, \mathbb{E}\big[Z^{\boldsymbol{g}_\ell} Z^{\bar{\boldsymbol{g}}_\ell}\big]\,,$$

where the convergence follows from the law of large numbers intuition via the Master Theorem. Consequently, the term $\sqrt{\frac{\tau}{n}}\, \boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell$ contains the correlation-driving factor

$$\tfrac{1}{n}\, \boldsymbol{W}_\ell^\top \boldsymbol{W}_\ell\, \boldsymbol{h}_{\ell-1},$$

whose i-th coordinate decomposes into

$$\frac{1}{n} \sum_j \boldsymbol{W}_{\ell,ji}^2\, \boldsymbol{h}_{\ell-1,i} \;+\; \frac{1}{n} \sum_j \sum_{k \neq i} \boldsymbol{W}_{\ell,ji} \boldsymbol{W}_{\ell,jk}\, \boldsymbol{h}_{\ell-1,k}.$$

By the law of large numbers, the first term converges to $Z^{\boldsymbol{h}_{\ell-1}}$, while the second behaves like a CLT term and is absorbed into a Gaussian field. Hence,

$$\left[\sqrt{\tfrac{\tau}{n}}\, \boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell\right]_i \xrightarrow{\text{a.s.}} \sqrt{\tau}\, \hat{Z}^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell} \;+\; \tau^2\, \mathring{b}_{\ell+1}\, Z^{\boldsymbol{h}_{\ell-1}},$$

where $\hat{Z}^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell}$ is centered Gaussian and correlated with $Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}$ via

$$\operatorname{Cov}\big(\hat{Z}^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell},\, Z^{\boldsymbol{W}_\ell^\top \boldsymbol{g}_\ell}\big) \;=\; \mathbb{E}\big[Z^{\bar{\boldsymbol{g}}_\ell} Z^{\boldsymbol{g}_\ell}\big]\,.$$

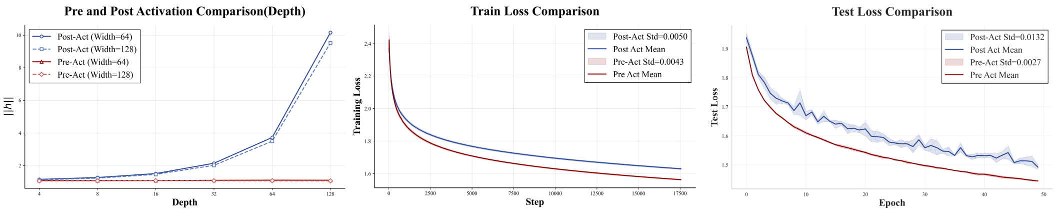
Collecting terms (still with $\phi = \mathrm{id}$), the coupled mean-field recursion becomes

$$\begin{aligned} Z^{\bar{\boldsymbol{g}}_L} &= Z^{\boldsymbol{g}_L} - \eta_c \mathcal{L}'(\mathring{f}, y)\, Z^{\boldsymbol{h}_L}, \\ Z^{\bar{\boldsymbol{g}}_{\ell-1}} &= Z^{\bar{\boldsymbol{g}}_\ell} - \tau \eta_c \mathcal{L}'(\mathring{f}, y)\, \mathbb{E}\big[Z^{\boldsymbol{g}_\ell} Z^{\bar{\boldsymbol{g}}_\ell}\big]\; Z^{\boldsymbol{h}_{\ell-1}} \\ &\qquad + \sqrt{\tau}\, Z^{\boldsymbol{W}_\ell^\top \bar{\boldsymbol{g}}_\ell} \;-\; \tau^2 \eta_c \mathcal{L}'(\mathring{f}, y)\, \mathbb{E}\big[Z^{\boldsymbol{g}_\ell} Z^{\bar{\boldsymbol{g}}_\ell}\big]\; Z^{\boldsymbol{h}_{\ell-1}}. \end{aligned}$$

Remark 2. *The last (higher-order) correction arises from reusing $\boldsymbol{W}_\ell$ and $\boldsymbol{W}_\ell^\top$ in both the forward and backward passes, and scales as τ^2 due to depth-adaptive normalization in ResNets. By Euler–Maruyama convergence, this term vanishes as $L \to \infty$ (with $\tau = T/L$), assuming the remaining quantities are well behaved. Moreover, the intuition developed here for the second forward and backward passes extends directly to any k-th pass.*

G Experiments

We now provide empirical evidence to validate and extend our theoretical results. Following the preliminary setup in Section 2, we conduct experiments on ResNets trained with SGD on CIFAR-10 under μP scaling and depth-adaptive normalization.



(a) Feature Norm across Depth (b) Training Loss (c) Test Loss

Figure 1: Comparison of pre-activation and post-activation ResNets: Pre-activation variant maintains stable feature norms across depth, while post-activation exhibits rapid growth. This stability leads to faster and more consistent training convergence, and results in lower test loss with reduced variance, indicating better generalization.

Comparison of Pre- and Post-Activation The empirical comparison between pre-activation and post-activation ResNets supports the result established in Proposition 1. As shown in Figure 1(a), post-activation ResNets exhibit rapid growth of feature norms as depth increases, while pre-activation networks maintain stable representations across all depths and widths. This stability translates into clear optimization advantages: in Figure 1(b), pre-activation networks converge faster and more consistently during training. As a result, the test loss in Figure 1(c) decreases more steadily and achieves lower values with reduced variance across runs.

Convergence to the Limiting FLDS. As established in Theorem 1, the feature evolution of ResNets trained with SGD converges to the limiting FLDS in the joint infinite-width–depth limit. To verify this result, we conduct experiments and report the outcomes in Figure 2 at initialization and after 10, 30, and 50 epochs of training. The first row of Figure 2 shows that the approximation error (MSE) decays as $O(1/L)$ with depth and as $O(1/n)$ with width, and that the two limits commute, directly echoing the convergence pattern predicted by Propositions 2 and 3. Moreover, the subsequent rows demonstrate that the same convergence pattern and rates observed at initialization persist throughout training. While Theorem 1 establishes convergence under the special order of taking width to infinity before depth and without an explicit rate, our experiments extend this result by empirically confirming both the explicit rates and their commutativity during training. Together, these findings strengthen Theorem 1 by showing not only convergence, but also the robustness of convergence rates under μP scaling.

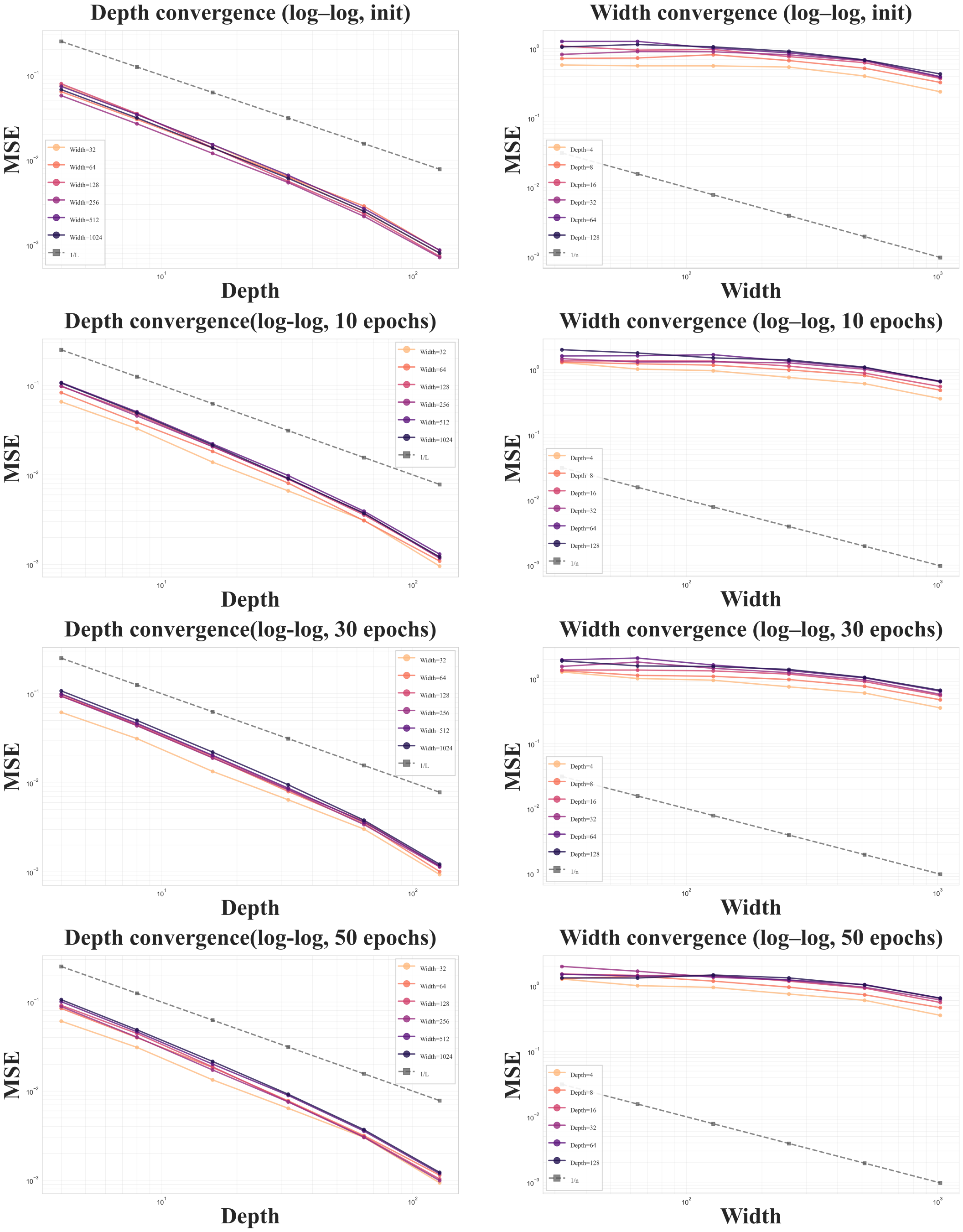


Figure 2: Depth and width convergence of ResNets under μP scaling at initialization and after 10, 30, and 50 epochs of training. In all cases, the approximation error (MSE) decays as $O(1/L)$ with depth and $O(1/n)$ with width, uniformly across the other factor, confirming commutativity of limits and robustness of convergence to the limiting FLDS.

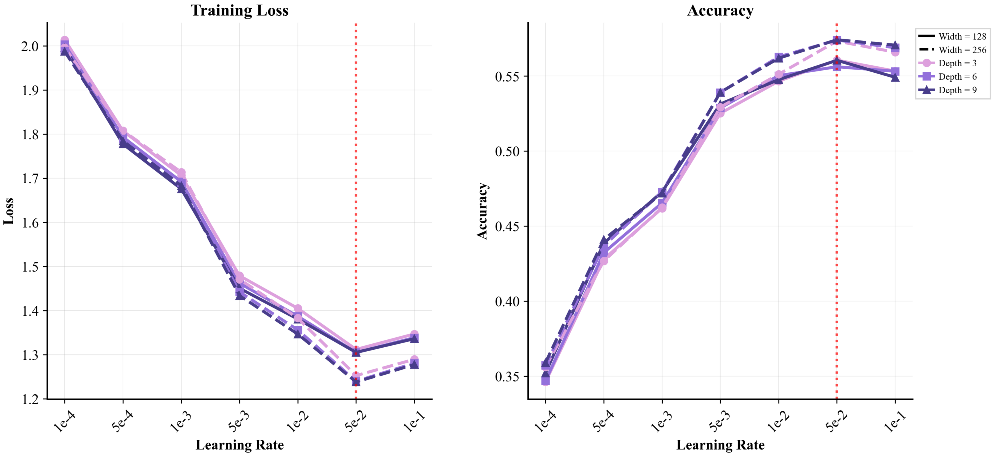


Figure 3: Hyperparameter transfer across depth and width under μP and depth-adaptive scaling. Training loss (left) and test accuracy (right) as a function of learning rate for MLPs with varying depth $(3, 6, 9)$ and width $(128, 256)$. The vertical red line marks the optimal learning rate selected at small scale. The same learning rate generalizes across depths and widths, confirming that μP scaling and depth-adaptive scaling together enable consistent performance without additional tuning.

Hyperparameter Transfer. An essential property of μP scaling is that hyperparameters tuned at small models can be transferred reliably to larger ones. Figure 3 illustrates this phenomenon for learning rate selection. On the left, the training loss decreases smoothly as the learning rate increases, with all depth–width configurations exhibiting the same qualitative trend. On the right, accuracy peaks near the same learning rate across depths $(3, 6, 9)$ and widths $(128, 256)$, as indicated by the red dashed line. The alignment of both loss reduction and accuracy gain demonstrates that the optimal learning rate identified at small models transfers directly to larger ones. This transferability is enabled jointly by μP scaling, which stabilizes training across widths, and by depth-adaptive scaling, which normalizes feature propagation across depths. Together, they preserve optimization dynamics across scales, allowing efficient hyperparameter tuning without the need for expensive re-optimization at larger sizes.