ON THE DEPTH TRANSFER OF NEURAL OPTIMIZERS VIA L-SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling deep learning models efficiently requires predictable performance transfer across compute regimes. Establishing reliable hyperparameter transfer allows optimal training dynamics and structural choices to be discovered on low-cost proxy models, avoiding catastrophic failures during massive-scale training. Existing depth-transfer analyses are architecture-specific and heuristic: hyperparameter transfer is only justified through toy blocks, and both rely on infinite-depth dynamics, a specific initialization, and feature-learning criteria. Building on L-smoothness framework of Xu et al. (2026), we understand hyperparameter transfer as how do the Lipschitz and L-smoothness constants of the loss scale with the depth K? We show that inverse-depth residual branch scaling $\alpha = \Theta(1/K)$ is sufficient to make both constants depth-independent at fixed width, uniformly over the bounded parameter domain and for arbitrary $p \to q$ geometries. Because the bounds depend on each residual branch only through its Lipschitz and smoothness constants, the argument applies to general residual networks—including branches of block depth ≥ 2 such as transformer and ResNet blocks—without assumptions on initialization or correlations between layer weights. Since the smoothness constant controls the curvature penalty in the descent inequality, our bounds motivate a learning-rate scale without an additional depth-dependent curvature correction for updates bounded in the prescribed norm. This provides a geometric explanation of inverse-depth residual scaling and its application to the operator-norm geometries associated with AdamW, Muon, and MOGA. Learning-rate sweeps on GPT-2 (depths 2–128) and CIFAR-100 ResNets (8–32 residual blocks) confirm that Muon and MOGA exhibit stable learning-rate transfer under these prescriptions, with a common optimal rate across a 64-fold increase in depth.

1 INTRODUCTION

Hyperparameter choices play a central role in the stability, efficiency, and performance of neural network training. Finding suitable values often requires many training runs, making hyperparameter tuning increasingly expensive as models grow (Yang et al., 2022; 2023a;b; Dey et al., 2025; Xu et al., 2026; Mlodozeniec et al., 2026; Jiang et al., 2026; Ma et al., 2026; Shigida et al., 2026). Moreover, settings that work well for a small model may become ineffective at a larger scale, limiting the value of inexpensive preliminary experiments. Understanding how hyperparameters should change with model size is therefore essential for translating insights from small experiments into reliable and efficient training of large models. In practice, many state-of-the-art optimizers show optimal learning rate drifts as depth scales up (Figure 1). In this paper, we aim to answer :

How can we design optimization algorithms so that optimal learning rates transfer reliably as network depth increases?

Prior work has approached depthwise hyperparameter transfer through parameterizations that preserve feature learning as network depth increases. Depth μP (Yang et al., 2023b) analyze training dynamics as width tends to infinity followed by depth, identifying feature learning and feature diversity as central design criteria for residual networks with one layer per branch. CompleteP (Dey et al., 2025) introduce complete feature learning to motivate residual scaling inversely proportional to depth and corresponding learning rate prescriptions for architectures with multiple layers per

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

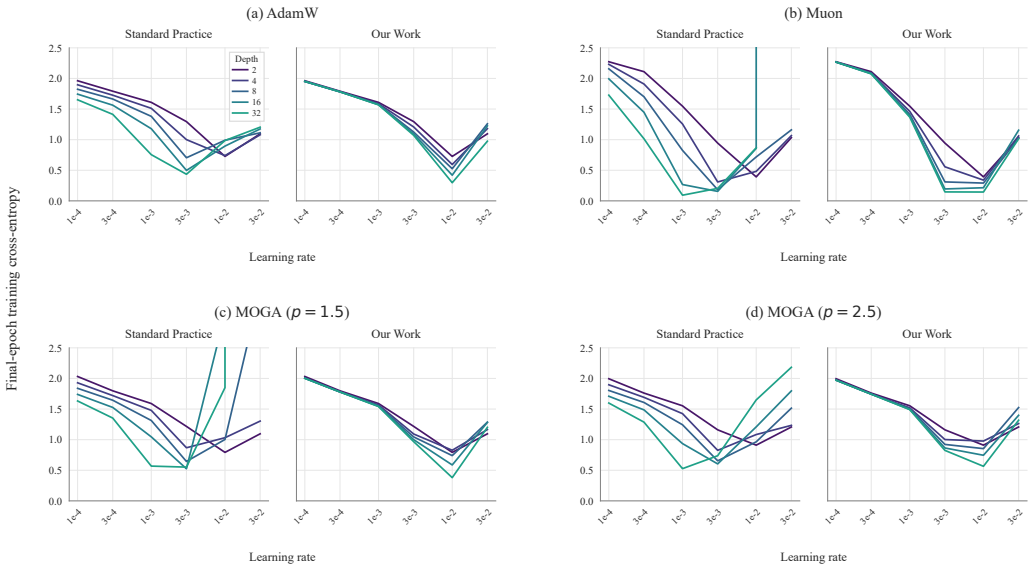

Figure 1: Final-epoch training loss against learning rate for residual MLPs of varying depth trained on CIFAR-10 with AdamW, Muon, and MOGA. Each branch contains two linear layers, and width is fixed across depths. Under standard unscaled residual branches (left), the best learning rate varies with depth. Our inverse-depth scaling, ($\alpha = K_{\text{base}}/K$) with ($K_{\text{base}} = 2$) (right), show that optimal learning rate is aligned across depths.

branch, including transformers. These theoretical analyses rely on specific architectural and initialization assumptions, are tailored to particular optimizers, and primarily characterize feature behavior in asymptotic regimes. However, preserving feature learning does not by itself establish that the optimal learning rate remains unchanged across depths. This raises a broader question: can we establish a general theory of depthwise learning rate transfer that applies to a broader class of architectures and optimizers, without relying on specific random initialization schemes or infinite width and depth limits?

To understand depthwise learning-rate transfer, we follow Xu et al. (2026) and characterize the stable learning rate through a scale-independent L-smoothness control. The L smoothness condition controls how much the loss gradient changes as the network weights vary, measuring weight perturbations in a norm adapted to the optimizer and gradient changes in its dual. L-smoothness governs step-size selection: a gradient step is guaranteed to decrease the loss only when the learning rate is of order $1/L$, and larger steps overshoot the region where the linearization remains valid. Consequently, the stable learning rate scales inversely with L. The obstacle to a uniform depth bound L is that the influence of an optimization step is not local. A full-network gradient update needs to perturb all layers simultaneously, requiring control of both perturbation propagation and second-order interactions between layers. **Our framework extends the depth scaling analysis beyond the Adam prescriptions developed in Depth μP and CompleteP to a broader class of optimizers used in LLM pretraining, including Muon and row normalization.** Our bounds apply directly to finite networks and do not require the sequential limit used in Depth μP and CompleteP, where the width is taken to infinity before depth.

Numerical results support our theoretical observations. We evaluate learning-rate transfer under this prescription through sweeps on GPT-2 language models and CIFAR-100 ResNets. The training loss sweeps show an aligned low-loss learning-rate region for Muon and MOGA across GPT-2 depths from 2 to 128. On CIFAR-100, all three evaluated MOGA settings share the same best tested learning rate across ResNets with 8, 16, and 32 residual blocks, measured by final-epoch training loss. AdamW also retains a common low-loss region, although its best tested rate shifts toward smaller values at the largest GPT-2 depths. Together, the theoretical bounds and empirical results provide a geometric account of depth scaling and support learning-rate transfer across the evaluated architectures and optimizer geometries.

2 PRELIMINARIES

2.1 SETTINGS, NOTATIONS AND ASSUMPTIONS

We consider the optimization problem of minimizing a neural network loss function:

$$\min_{\theta} \mathcal{L}(\theta) := l(y_K(x;\theta))$$

Let $(H, \|\cdot\|_H)$ be the hidden-state space and let $(P, \|\cdot\|_P)$ be the parameter space. Then for all k, we have $y_k \in H$, $W_k, U_k, V_k \in P$, and $f_k : H \times P \to H$. Throughout this paper, we assume that the branch function f_k is twice continuously Fréchet differentiable. The derivatives are maps $\partial_y f_k : H \to H$ and $\partial_{W_k} f_k : P \to H$. The second derivatives are maps $\partial_y^2 f_k : H \times H \to H$, $\partial_y \partial_{W_k} f_k : H \times P \to H$, and $\partial_{W_k}^2 f_k : P \times P \to H$.

For a parameter tuple $W = (W_1, \ldots, W_K)$, define the block norm

$$\|W\|_{\text{block}} = \max_{1 \le k \le K} \|W_k\|_P ,$$

and its dual

$$\|G\|_{\text{block},*} = \sum_{k=1}^{K} \|G_k\|_{P^*} ,$$

where $\|\cdot\|_{P^*}$ denotes the dual norm on P^*.

For notational simplicity, we write $f_{W_k}(y) := f_k(y, W_k)$.

Definition 1 (Abstract Residual Network)**.** Let $(H, \|\cdot\|_H)$ and $(P, \|\cdot\|_P)$ be normed vector spaces. Let $x \in H$ be the input. Define $y_0 = x$ and for $k = 1, \ldots, K$,

$$y_k = y_{k-1} + \alpha f_{W_k}(y_{k-1}),$$

where $f_{W_k} : H \to H$ is a twice continuously differentiable residual function parameterized by $W_k \in P$. In this paper, we set the scaling factor as

$$\alpha = \frac{\beta}{K}, \qquad \beta = \Theta(1).$$

The network output is $y_K \in H$, and the loss is

$$\mathcal{L}(W) = l(y_K),$$

with $l : H \to \mathbb{R}$. The parameter tuple is $W = (W_1, \ldots, W_K)$, and the perturbation directions are $\Delta U = (\Delta U_1, \ldots, \Delta U_K)$ and $\Delta V = (\Delta V_1, \ldots, \Delta V_K)$, with $\Delta U_k, \Delta V_k \in P$.

We make the following regularity assumptions on the network we defined. They are imposed throughout the paper.

Assumption 1 (Initial Condition and Perturbation Bounds)**.**

$$\|x\|_H \le C_x, \qquad \|\Delta U_k\|_P \le 1, \qquad \|\Delta V_k\|_P \le 1.$$

Assumption 2 (Residual Function Regularity)**.** There exist positive constants C_f, L_f, M_f, C_W, M_{Wf}, and M_{WW}, independent of K, such that the following bounds hold for every $k = 1, \ldots, K$, every $y, y_1, y_2 \in H$, every $W_k, W_k' \in P$, and every $\Delta U_k, \Delta V_k \in P$. In appendix C.1 and C.2, with mild assumptions on parameter space and function regularity, we can satisfy our abstract residual function regularity for MLP and Attention blocks

2.1 **First Order Regularity in Hidden State**

$$\|f_{W_k}(y_1) - f_{W_k}(y_2)\|_H \le L_f \|y_1 - y_2\|_H$$

2.2 **Second Order Regularity in the Hidden State**

$$\|\partial_y f_{W_k}(y_1) - \partial_y f_{W_k}(y_2)\|_{H \to H} \le M_f \|y_1 - y_2\|_H$$

2.3 **First Order Regularity in the Parameters**

$$\left\| f_{W_k}(y) - f_{W_k'}(y) \right\|_H \leq C_W \left\| W_k - W_k' \right\|_P$$

2.4 **Mixed-State Parameter Regularity**

$$\left\| \partial_y f_{W_k}(y) - \partial_y f_{W_k'}(y) \right\|_{H \to H} \leq M_{Wf} \left\| W_k - W_k' \right\|_P$$

2.5 **Second Order Regularity in the Parameters**

$$\left\| \partial_{W_k} f_{W_k}(y)[\Delta U_k] - \partial_{W_k'} f_{W_k'}(y)[\Delta U_k] \right\|_H \leq M_{WW} \left\| W_k - W_k' \right\|_P \left\| \Delta U_k \right\|_P$$

2.6 **Bounded Output at Initial State**

$$\left\| f_{W_k}(y_0) \right\|_H \leq C_f$$

Remark 1. Following (Xu et al., 2026), we show that our assumption is realistic. Their analysis was conducted on MLP and attention blocks, and they imposed activation function regularity and bounded parameter assumptions that are consistent with our generalized abstract-residual function Assumption 2 for the corresponding MLP and Attention blocks. Our set-up is therefore a faithful abstraction of the standard conditions satisfied by modern architectures; the detailed reductions are carried out in the appendix via Proposition 1 and Proposition 2.

Assumption 3 (Regularity of the Loss Function)**.** There exist constants $L_J, M_J > 0$, independent of K, such that for all $z \in H$,

$$\left\| \nabla l(z) \right\|_{H^*} \leq L_J, \qquad \left| \nabla^2 l(z)[u, v] \right| \leq M_J \left\| u \right\|_H \left\| v \right\|_H .$$

2.2 Modern Neural Optimizers

Optimizers as Steepest Descent Under Matrix Operator Norms Many state-of-the-art optimizers can be approximated by steepest descent under some matrix operator norm $\|\cdot\|_{p\to q}$ (Bernstein & Newhouse, 2024b; Pethick et al., 2025; Xu et al., 2026). The steepest descent direction D^* is given by

$$D^* = \underset{\|D\|_{p\to q}\leq 1}{\arg\min} \langle \nabla f(\theta), D \rangle, \text{ where } \|D\|_{p\to q} := \sup_{x\neq 0} \frac{\|Dx\|_q}{\|x\|_p} = \sup_{\|x\|_p=1} \|Dx\|_q$$

For example, we will recover Adam (Kingma & Ba, 2015; Balles & Hennig, 2018), SignSGD (Bernstein et al., 2018) and Lion (Chen et al., 2023) as steepest descent under $\ell_1 \to \ell_\infty$, MUON as steepest descent under $\ell_2 \to \ell_2$ (Jordan et al., 2024; Bernstein & Newhouse, 2024b) and Row normalization as steepest descent under $\ell_2 \to \ell_\infty$ (Wen et al., 2025b; Yuan et al., 2026; Deng et al., 2026; Xu et al., 2026) .

Steepest descent under matrix operator norms is well motivated because these norms capture the smoothness structure of neural networks: with respect to a suitable operator norm on each layer, the network output and loss are Lipschitz and smooth in the weights with constants that remain controlled as the architecture scales (Xu et al., 2026), so the resulting steepest-descent update, obtained by dualizing the gradient under this norm, takes steps adapted to the true local geometry of the loss (Large et al., 2024; Bernstein & Newhouse, 2024a; Xu et al., 2026).

Hyperparameter Transfer. Hyperparameter transfer, specifically learning-rate transfer, is the technique of reusing a learning rate tuned on a smaller network for the training of a larger model which eliminates the cost of redundant hyperparameter tuning. In particular, we study learning rate transfer over depth. Depth-μP (Yang et al., 2023b) uses the tensor programs framework Yang (2020); Yang et al. (2022) to study the infinite-depth limit of training dynamics and derive parameterization rules that satisfy the activation and gradient scales feature learning rules as depth varies. However, depth-μP holds only for block depth 1 models and their scaling rules deteriorate for the commonly used transformer, which has block depth 2. To fix the limitation of Depth-μP, Complete-P (Dey et al., 2025) takes the μP framework and derives another set of rules targeted at block-depth ≥ 2

architectures. Both works are limited to block-depth constraints and rely heavily on infinite-depth initialization and early training dynamics. Furthermore, rigorous convergence analysis of training dynamics are only established for block-depth 1, meaning that feature learning analysis cannot solely establish the connection to how the network evolves under training for block-depth ≥ 2 architecture. In contrast, our analysis is robust to block-depth; in particular, we provide a worst-case guarantee that is sufficient for architectures of any block-depth. Our analysis also directly controls the gradient scale and loss sensitivity of the network as depth increases which provides a more principled alternative explanation to learning rate transfer scaling rules for transformers. Lastly, our approach does not require specific initialization configurations or weight correlation assumptions, providing learning-rate scaling rules beyond specialized cases in prior works.

Zheng et al. (2026) develop spectral conditions for joint width-depth scaling across residual architectures and optimizers, including Muon. Their parameterizations include inverse-depth residual scaling and learning rates without a depth correction at fixed width. Our analysis provides a complementary, unified framework through depth-uniform bounds on loss sensitivity and curvature under prescribed parameter norms. By expressing the bounds in terms of common residual-function regularity conditions, the same argument applies across architectures and norm geometries satisfying these conditions.

Norm Geometry and L-smoothness. Large et al. (2024) establish Lipschitz and smoothness guarantees in a modular norm constructed recursively from module sensitivities and mass allocations, including depth-independent bounds for a suitably scaled residual construction (Propositions 4–5). We study depth scaling directly in a prescribed parameter geometry. For the identity-skip recurrence $y_k = y_{k-1} + (\beta/K) f_k(y_{k-1})$, we establish depth-uniform Lipschitz and smoothness bounds at fixed width in the unweighted maximum of prescribed parameter-block norms, under uniform branch and loss regularity assumptions. Specializing these norms to $p \to q$ operator norms makes the associated unit-ball steepest-descent directions explicit and computable blockwise, without introducing separate module-mass parameters. The resulting smoothness bound gives a depth-independent curvature penalty for updates measured in this geometry, providing the basis for the learning-rate discussion in Section 3.2 and our optimizer-specific prescriptions. The analysis treats each residual branch through its regularity constants and therefore also covers multi-layer branches satisfying these assumptions.

3 Main Theorems

3.1 Depth-Independent Lipschitz Constant

Xu et al. (2026) showed that Lipschitz continuity can characterize the sensitivity of network loss to parameter perturbations and established a width-independent Lipschitz bound using mean-normalized norm, making the network stable in width. Here, we study whether an appropriate branch scaling can yield a depth-independent Lipschitz bound to make the loss function sensitivity controlled in depth. Denote $\delta y_a = \nabla_{W_a} y_i[\Delta U_a]$ the first order effect of perturbing W_a in the direction ΔU_a. For layers after a, the residual Jacobian $J_i = \partial y_i / \partial y_{i-1}$ maps the incoming perturbation to the next layers. The chain rule gives:

$$\delta y_i = \delta y_{i-1} + \alpha \partial_y f_{W_i}(y_{i-1})[\delta y_{i-1}] = J_i \delta y_{i-1}$$

Therefore for a later layer b we have

$$\delta y_b = J_b \dots J_{a+1} \delta y_a$$

Under our assumptions, we have the following upper bound for the Jacobians (see the Appendix for details)

$$\|J_i\|_{H \to H} \leq 1 + C\alpha$$

Where $C \in O(1)$. Therefore, we have

$$\|\delta y_b\|_H \leq (1 + C\alpha)^{b-a} \|\delta y_a\|_H \leq (1 + C\alpha)^K \|\delta y_a\|_H$$

Which shows that the worst case guarantee blows up as $K \to \infty$ (number of layers between a and b increases). However, an appropriate branch scaling can control this bound. Setting $\alpha = 1/K$, we

have

$$\|\delta y_b\|_H \le (1+\frac{C}{K})^{b-a}\|\delta y_a\|_H \le (1+\frac{C}{K})^K\|\delta y_a\|_H \le \exp(C)\|\delta y_a\|_H$$

Theorem 1 (Depth-Independent Lipschitz Continuity)**.** *Under Assumptions 1 to 3 and $\alpha = \beta/K$, $\beta = \Theta(1)$, there exists $L_{\mathrm{Lip}} = O(1)$, independent of K, such that*

$$\sup_{\|\Delta U\|_{\mathrm{block}} \le 1} |\nabla \mathcal{L}(W)[\Delta U]| \le L_{\mathrm{Lip}}.$$

Consequently,

$$|\mathcal{L}(W^1) - \mathcal{L}(W^2)| \le L_{\mathrm{Lip}}\|W^1 - W^2\|_{\mathrm{block}}.$$

3.2 Depth-Independent L-smoothness Theorem

While Lipschitz continuity is enough to control the sensitivity of the network to weight perturbations, it provides limited information on the actual optimization dynamics during training. We show that inverse depth branch scaling provides a depth-independent L-smoothness constant that directly controls how much gradients can vary along the optimization trajectory. Bounding this quantity is crucial, as it ensures that the steps we take in optimization are stable as depth varies.

Theorem 2 (Depth-Independent L-Smoothness)**.** *Under Assumptions 1 to 3 and $\alpha = \beta/K$, $\beta = \Theta(1)$, there exists $L = O(1)$, independent of K, such that*

$$\sup_{\|\Delta U\|_{\mathrm{block}} \le 1, \|\Delta V\|_{\mathrm{block}} \le 1} |\nabla^2 \mathcal{L}(W)[\Delta U, \Delta V]| \le L.$$

Consequently,

$$\|\nabla \mathcal{L}(W^1) - \nabla \mathcal{L}(W^2)\|_{\mathrm{block},*} \le L\|W^1 - W^2\|_{\mathrm{block}}.$$

Why L-Smoothness Motivates Learning Rate Transfer Following Xu et al. (2026), we use the smoothness constant L to guide learning rate selection. For an update $x^+ = x + \eta D$ along a unit norm steepest descent direction, the linear approximation predicts a loss decrease of $\eta\|\nabla f(x)\|_*$, while the curvature penalty is bounded by $L\eta^2/2$. Minimizing the resulting upper bound on the loss gives $\eta^* = \|\nabla f(x)\|_*/L$. A common smoothness bound across depths therefore ensures that the curvature term does not require smaller learning rates as the network becomes deeper. If the dual gradient norm also remains comparable across depths, this rule yields learning rates of comparable magnitude, providing a theoretical motivation for depthwise learning rate transfer.

3.3 Proof Sketch

We outline the idea behind Theorem 1 and Theorem 2. All constants below are independent of the depth K; the whole argument rests on a $1/K$ cancellation against a K-fold summation.

(i) A stable forward pass. With $\alpha = \beta/K$, each residual step $y_k = y_{k-1} + \alpha f_{W_k}(y_{k-1})$ perturbs the state by only $O(1/K)$. Since there are only K steps, the states remain uniformly bounded, $\|y_k\|_H \le C_y$ for all k (Corollary 1). Hence every quantity evaluated at the iterates y_{k-1} is bounded a priori, independently of K.

(ii) Controlled Jacobian products. The layer Jacobian $J_k = I + \alpha\,\partial_y f_{W_k}(y_{k-1})$ satisfies $\|J_k\|_{H\to H} \le 1 + \alpha L_f$, so arbitrary products obey

$$\|J_K \cdots J_j\|_{H\to H} \le (1+\beta L_f/K)^K \le e^{\beta L_f} = C_J = O(1)$$

(Corollary 3). Perturbations thus don't blow up as they propagate through the network.

(iii) First-order directional derivatives are $O(1/K)$ each. By the chain rule, $\nabla_W \mathcal{L}[\Delta U] = l'(y_K)\sum_j s_K^{(j,U_j)}$. Each term $s_K^{(j,U_j)}$ originates from the single source $\alpha\,\partial_{W_j} f_{W_j}(y_{j-1})[\Delta U_j]$ at layer j and is then propagated by $J_K \cdots J_{j+1}$, giving

$$\left\|s_K^{(j,U_j)}\right\|_H \le C_J \cdot \alpha C_W = O(1/K).$$

Summing over the K layers, $\left\|\sum_j s_K^{(j,U_j)}\right\| \le C_s = O(1)$ (Corollary 5).

(iv) Second-order directional derivatives: $O(1/K)$ diagonal, $O(1/K^2)$ off-diagonal. The Hessian splits into an outer-product part $l''(y_K)(\sum_j s_K^{(j)})(\sum_\ell s_K^{(\ell)}) = O(1)$ and a propagation part $l'(y_K)\sum_{j,\ell} h_K^{(j,\ell)}$. In the latter, the source carries two derivative hits, hence an α^2 factor:

$$\left\|h_K^{(j,j)}\right\|_H = O(1/K), \qquad \left\|h_K^{(j,\ell)}\right\|_H = O(1/K^2) \quad (j \neq \ell).$$

There are K diagonal pairs and $K(K-1)$ off-diagonal pairs, so

$$\left\|\sum_{j,\ell} h_K^{(j,\ell)}\right\| \leq K \cdot O(1/K) + K^2 \cdot O(1/K^2) = O(1).$$

(v) From directional bounds to Lipschitz/smoothness. Having shown $\sup_{\|\Delta U\|\leq 1} |\nabla \mathcal{L}[\Delta U]|$ and $\sup_{\|\Delta U\|,\|\Delta V\|\leq 1} |\nabla^2 \mathcal{L}[\Delta U, \Delta V]|$ are uniformly bounded, Lemma 1 and Lemma 2 (via the fundamental theorem of calculus) upgrade these to the Lipschitz and smoothness statements in the block norm, completing the proof of Theorem 1 and Theorem 2.

4 Experiments

We evaluate our scaling laws on MOGA, Muon, and Adam. To examine the robustness of MOGA, we run on the norm parameter $p \in \{1.5, 2, 2.5\}$. We examine large-scale pretraining tasks on GPT2 (Radford et al., 2019) and the image classification tasks on ResNet (He et al., 2016) to show that our scaling prescriptions are robust to architecture.

4.1 Learning Rate Transfer On Transformers

We use the GPT2 Model (Radford et al., 2019) and equip it with the standard tokenizer of vocabulary size 50257, context-length 1024, and the OpenWebText Dataset (Gokaslan & Cohen, 2019). Below are the detailed configurations:

1. **Model** We run GPT2 with fixed 256 width and 4 attention heads and vary the depth of the model over $\{2, 4, 8, 16, 32, 64, 128\}$. This model is run with Muon, AdamW, and Moga $p \in \{1.5, 2, 2.5\}$.
2. **Learning-rate Schedule and Training Configuration** We run GPT2 from scratch with 256 effective batch size and 1200 iterations, corresponding to a 315 Million token budget per run. We use cosine learning rate decay with 10% warmup and min_lr = 0.1 peak_lr meaning that the learning rate will increase linearly to the peak learning rate and decay with a cosine schedule to 10% of the peak learning rate at the end of training. We sweep across six learning rates $\{1 \times 10^{-3}, 2 \times 10^{-3}, 4 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 4 \times 10^{-2}\}$.

As shown in figure 2, models spanning a 64-fold depth counts choose nearly the same optimal learning rate. This implies that the optimal learning rate is effectively invariant to model depth for MOGA and Muon. We observe that optimal learning rates tuned for shallow models can generally be transferred directly to deeper models, thus eliminating the cost of retuning hyperparameters.

As our theory predicted, choosing different p values for MOGA does not affect its learning rate transfer as none of the optimizer geometries require additional learning rate corrections. Of all the optimizers, Muon and MOGA exhibit a wide stable learning rate area that transfers as depth increases, indicating strong stable learning rate transfer. Notably, Adam becomes more unstable at large depths but the optimal learning rate shifts to the left only by a factor of 2; which still shows effective learning rate transfer.

4.2 Learning Rate Transfer On ResNet

We use the standard ResNet architecture (He et al., 2016) and equip it with the CIFAR100 dataset(Krizhevsky, 2009). Below are the detailed configurations:

1. **Model** We train on CIFAR-adapted ResNets with standard two-convolution BasicBlocks. We use stage configurations of $[2, 2, 2, 2]$,$[3, 4, 6, 3]$ and $[8, 8, 8, 8]$ corresponding to

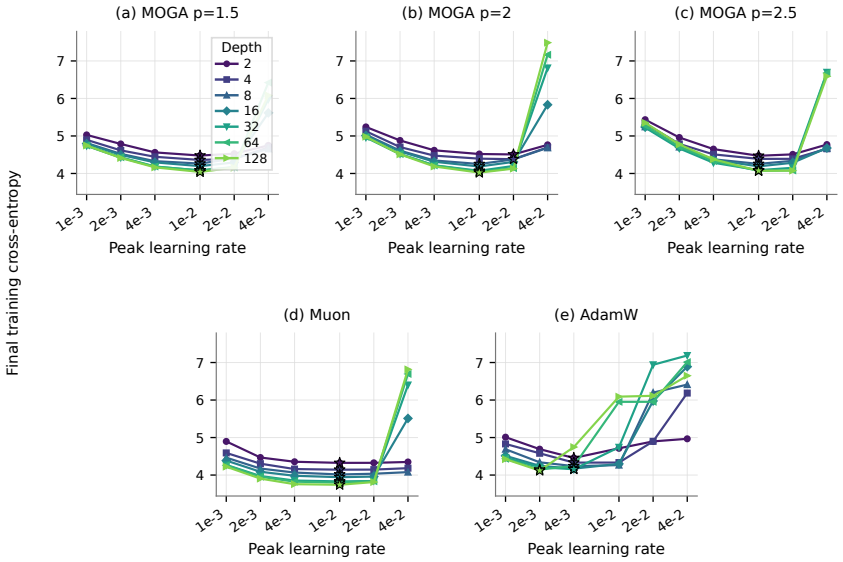

Figure 2: Learning-rate transfer across GPT-2 depths at fixed width. Panels show MOGA with $p \in \{1.5, 2, 2.5\}$, Muon, and AdamW. Colors and markers indicate transformer depth. Stars mark the lowest training loss among the tested learning rates at each depth.

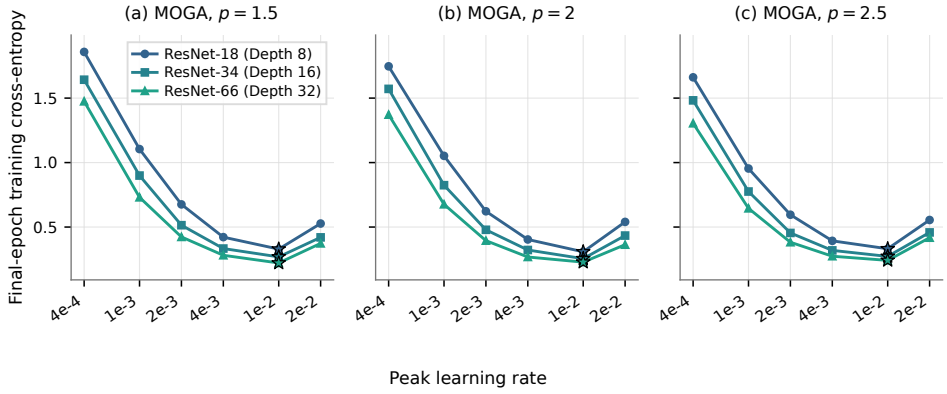

Figure 3: Learning-rate transfer on CIFAR-100 with depth-scaled ResNets. Panels show MOGA with $p \in \{1.5, 2, 2.5\}$. Curves report final-epoch training cross-entropy for ResNet-18, ResNet-34, and ResNet-66, corresponding to $K \in \{8, 16, 32\}$ residual blocks. Stars indicate the lowest loss among the tested learning rates for each model.

ResNet18, ResNet34, and ResNet66 with depths of $K \in \{8, 16, 32\}$ residual blocks respectively. We train on the MOGA optimizer with $p \in \{1.5, 2, 2.5\}$. We train on CIFAR-100, splitting its 50,000 training examples into 45,000 training examples and 5,000 validation examples using a fixed split seed. Training images are augmented using random 32×32 crops with four-pixel padding and random horizontal flipping. All images are normalized using the CIFAR-100 channel statistics.

2. **Learning-rate Schedule and Training Configuration**. We train every model with batch size 64 for 20 epochs, corresponding to 703 steps per epoch and a total of 14060 updates. We use one epoch of linear warm up and cosine decay to 50% of the peak learning rate. We sweep across six learning rates $\{4\times10^{-4}, 1\times10^{-3}, 2\times10^{-3}, 4\times10^{-3}, 1\times10^{-2}, 2\times10^{-2}\}$

In figure 3, the three selected models select the same optimal learning rate over all 3 variants of the MOGA optimizer. An optimal learning rate tuned on ResNet-18 can be directly transferred to ResNet-66. Another particularly strong result is that all three variants showed matching stable

learning rate areas that are also robust in depth. Again, these results support our claim that MOGA consistently transfers learning rates for different choices of p. These results also provide empirical evidence that our scaling prescriptions work for different architectures; in total we have tested GPT-2, ResNet, and residual MLPs for our baseline.

5 CONCLUSION

In this paper, we adopt the matrix-operator norm steepest descent framework from Xu et al. (2026) to investigate how optimizer geometry can yield depthwise learning rate scaling rules. We show that many state-of-the-art optimizers- such as AdamW and MUON- can be studied through associated matrix-operator norm geometries. However, our analysis shows that neural networks are not depth-independent under the standard branch scaling; the worst case Lipschitz bound becomes unstable as network depth increases.

We propose inverse-depth branch scaling to bound aggregate sensitivity to layer updates uniformly in depth and hence induce an L-Lipschitz constant that is independent of depth. Our further analysis of gradient sensitivity shows that the objective loss function under inverse-depth branch scaling sufficiently induces a depth-independent L-smooth constant which motivates a learning rate scale without an additional depth-dependent curvature correction. Our experiments evaluate the resulting learning-rate transfer for practical optimizers.

Our experiments of large-scale pre-training tasks on GPT-2 and image classification tasks on ResNet show that optimal learning rates stay nearly invariant on the tested state-of-the-art optimizers, which allows models over a 64-fold depth range to effectively transfer learning rates from small models to deeper models, reducing the computation budget of hyperparameter tuning.

Our analysis establishes inverse-depth residual scaling as a worst-case guarantee for depth-uniform sensitivity and smoothness bounds. Larger residual multipliers can also support learning-rate transfer in certain settings. For example, Depth-μP (Yang et al., 2023b) uses ($K^{-1/2}$) residual scaling and demonstrates transfer in residual networks with single-layer branches under its parameterization. Our present bounds do not explain this behavior. Extending the analysis to account for the architectural or statistical structure that enables transfer in these special settings is a natural direction for future work.

REFERENCES

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning, 2024a. URL https://arxiv.org/abs/2410.21265. arXiv:2410.21265.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024b. URL https://arxiv.org/abs/2409.20325. arXiv:2409.20325.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

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

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36:49205–49233, 2023.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: CompleteP enables compute-efficient deep transformers. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.01618v4.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. Memory-efficient LLM pretraining via minimalist optimizer design. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.16659.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://arxiv.org/abs/1512.03385.

Tianze Jiang, Blake Bordelon, Cengiz Pehlevan, and Boris Hanin. Hyperparameter transfer with mixture-of-expert layers. *arXiv preprint arXiv:2601.20205*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization, 2025. URL https://arxiv.org/abs/2503.12645. arXiv:2503.12645.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. *Advances in Neural Information Processing Systems*, 37: 73501–73548, 2024.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon, 2025. URL https://arxiv.org/abs/2502.02900. arXiv:2502.02900.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere, 2024. URL https://arxiv.org/abs/2410.01131. arXiv:2410.01131.

Yuxin Ma, Nan Chen, Mateo Díaz, Soufiane Hayou, Dmitriy Kunisky, and Soledad Villar. μpscaling small models: Principled warm starts and hyperparameter transfer. *arXiv preprint arXiv:2602.10545*, 2026.

Bruno Mlodozeniec, Pierre Ablin, Louis Béthune, Dan Busbridge, Michal Klein, Jason Ramapuram, and Marco Cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. In *International Conference on Learning Representations*, volume 2026, pp. 45300–45325, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs, 2025. URL https://arxiv.org/abs/2502.07529. arXiv:2502.07529.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs), 2025. URL https://arxiv.org/abs/2505.13416. arXiv:2505.13416.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon, 2025. URL https://arxiv.org/abs/2505.23737. arXiv:2505.23737.

Boris Shigida, Boris Hanin, and Andrey Gromov. Learning rate transfer in normalized transformers, 2026. URL https://arxiv.org/abs/2604.27077. arXiv:2604.27077.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them, 2025a. URL https://arxiv.org/abs/2509.02046. arXiv:2509.02046.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization, 2026. URL https://arxiv.org/abs/2606.16899. arXiv:2606.16899.

Ziqing Wen, Yanqi Shi, Jiahuan Wang, Ping Luo, Linbo Qiao, Dongsheng Li, and Tao Sun. Sron: State-free llm training via row-wise gradient normalization. 2025b.

Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms I: Row/column normalization and hyperparameter transfer, 2026. URL https://arxiv.org/abs/2603.09952v1. arXiv:2603.09952.

Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL https://arxiv.org/abs/2011.14522.

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

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning, 2023a. URL https://arxiv.org/abs/2310.17813. arXiv:2310.17813.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite-depth neural networks, 2023b. URL https://arxiv.org/abs/2310.02244. arXiv:2310.02244.

Jinghui Yuan, Jiaxuan Zou, Shuo Wang, Yong Liu, and Feiping Nie. Nora: Normalized orthogonal row alignment for scalable matrix optimizer. *arXiv preprint arXiv:2605.03769*, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1910.07467.

Chenyu Zheng, Rongzhen Wang, Xinyu Zhang, and Chongxuan Li. Spectral condition for μP under width-depth scaling, 2026. URL https://arxiv.org/abs/2603.00541. arXiv:2603.00541.

A LLM USAGE

We used LLMs to assist with manuscript drafting and editing, literature searches, reviewing and revising mathematical arguments, and debugging experimental code. The authors reviewed and verified the AI-assisted material included in the submission and take responsibility for the final manuscript, theoretical claims, code, and reported results.

B PROOF OF THE MAIN THEOREM

B.1 COROLLARIES

Corollary 1 (Forward Boundedness)**.** For all $k = 0, \dots, K$,

$$\|y_k\|_H \le C_y := e^{\beta L_f}\Big(C_x + \frac{L_f C_x + C_f}{L_f}\Big).$$

Proof. By Assumption 1 and Assumption 2,

$$\begin{aligned}\|f_{W_k}(y_{k-1})\|_H &\le \|f_{W_k}(y_{k-1}) - f_{W_k}(y_0)\|_H + \|f_{W_k}(y_0)\|_H \le L_f \|y_{k-1} - y_0\|_H + C_f \\ &\le L_f \|y_{k-1}\|_H + L_f C_x + C_f.\end{aligned}$$

Let $u_k := \|y_k\|_H$. Then

$$u_k \le (1 + \alpha L_f) u_{k-1} + \alpha(L_f C_x + C_f).$$

Unrolling,

$$u_k \le (1 + \alpha L_f)^k u_0 + \alpha(L_f C_x + C_f) \sum_{i=0}^{k-1} (1 + \alpha L_f)^i.$$

Since $u_0 \le C_x$ and $(1 + \alpha L_f)^k \le e^{\beta L_f}$,

$$u_k \le e^{\beta L_f}\Big(C_x + \frac{L_f C_x + C_f}{L_f}\Big).$$

□

Corollary 2 (Jacobian Bound)**.** For all $k = 1, \dots, K$, with $J_k := I + \alpha \partial_y f_{W_k}(y_{k-1})$,

$$\|J_k\|_{H \to H} \le 1 + \alpha L_f.$$

Proof. By Assumption 2 and the mean value theorem,

$$\|\partial_y f_{W_k}(y_{k-1})\|_{H \to H} \le L_f.$$

Hence

$$\|J_k\|_{H \to H} \le \|I\|_{H \to H} + \alpha \|\partial_y f_{W_k}(y_{k-1})\|_{H \to H} \le 1 + \alpha L_f.$$

□

Corollary 3 (Jacobian Product Bound)**.** For all $1 \le j \le k \le K$,

$$\Big\|J_k J_{k-1} \cdots J_j\Big\|_{H \to H} \le (1 + \alpha L_f)^{k-j+1} \le e^{\beta L_f} =: C_J.$$

Proof. By Corollary 2,

$$\Big\|J_k J_{k-1} \cdots J_j\Big\|_{H \to H} \le \prod_{i=j}^{k} \|J_i\|_{H \to H} \le (1 + \alpha L_f)^{k-j+1}.$$

Substituting $\alpha = \beta/K$ and using $(1 + x/K)^K \le e^x$ for $x \ge 0$,

$$(1 + \alpha L_f)^{k-j+1} \le (1 + \beta L_f / K)^K \le e^{\beta L_f}.$$

□

Corollary 4 (Derivative Bounds Implied by Assumption 2)**.** Under Assumption 2, the following bounds hold for all $k = 1, \dots, K$ and all $y \in H$:

$$\|\partial_y f_{W_k}(y)\|_{H\to H} \le L_f, \tag{1}$$
$$\left\|\partial_y^2 f_{W_k}(y)[u,v]\right\|_H \le M_f \|u\|_H \|v\|_H, \tag{2}$$
$$\|\partial_{W_k} f_{W_k}(y)\|_{P\to H} \le C_W, \tag{3}$$
$$\|\partial_y \partial_{W_k} f_{W_k}(y)[h,w]\|_H \le M_{Wf} \|h\|_H \|w\|_P, \tag{4}$$
$$\left\|\partial_{W_k}^2 f_{W_k}(y)[u,w]\right\|_H \le M_{WW} \|u\|_P \|w\|_P. \tag{5}$$

Proof. We prove equation 1 in detail; the remaining bounds follow by the same argument applied to the corresponding Lipschitz condition in Assumption 2.

By Assumption 2, f_{W_k} is L_f-Lipschitz in y:

$$\|f_{W_k}(y_1) - f_{W_k}(y_2)\|_H \le L_f \|y_1 - y_2\|_H, \qquad \forall y_1, y_2 \in H.$$

Fix $y \in H$ and $h \in H$. For any $t > 0$, set $y_1 = y + th$ and $y_2 = y$. Then

$$\|f_{W_k}(y + th) - f_{W_k}(y)\|_H \le L_f \|th\|_H = L_f\, t\, \|h\|_H.$$

Dividing by t gives

$$\left\|\frac{f_{W_k}(y + th) - f_{W_k}(y)}{t}\right\|_H \le L_f \|h\|_H.$$

Taking the limit as $t \to 0$ and using the definition of the directional derivative,

$$\|\partial_y f_{W_k}(y)[h]\|_H \le L_f \|h\|_H.$$

Taking the supremum over all $h \in H$ with $\|h\|_H \le 1$,

$$\|\partial_y f_{W_k}(y)\|_{H\to H} = \sup_{\|h\|_H \le 1} \|\partial_y f_{W_k}(y)[h]\|_H \le L_f.$$

This proves equation 1.

For equation 2, apply the same argument to the M_f-Lipschitz condition on $\partial_y f_{W_k}$. For equation 3, use the C_W-Lipschitz condition on f_{W_k} in W_k. For equation 4, use the M_{Wf}-Lipschitz condition on $\partial_y f_{W_k}$ in W_k. For equation 5, use the M_{WW}-Lipschitz condition on $\partial_{W_k} f_{W_k}[\Delta U_k]$ in W_k. □

B.2 First-Order Directional Derivative Bound

To simplify our notation, we define:

Definition 2 (Directional Derivatives)**.** For $k = 0, \dots, K$ and $j, \ell = 1, \dots, K$, define

$$s_k^{(j,U_j)} := \nabla_{W_j} y_k[\Delta U_j] \in H,$$

$$h_k^{(j,\ell;U_j,V_\ell)} := \nabla^2_{W_j,W_\ell} y_k[\Delta U_j, \Delta V_\ell] \in H.$$

Then for all $j = 1, \dots, K$, we have:

$$\left\|s_K^{(j,U_j)}\right\|_H \le \frac{C_s}{K}, \qquad C_s := C_J \beta C_W.$$

Consequently,

$$\left\|\sum_{j=1}^{K} s_K^{(j,U_j)}\right\|_H \le C_s.$$

Proof. Differentiating $y_k = y_{k-1} + \alpha f_{W_k}(y_{k-1})$ with respect to W_j in direction ΔU_j,

$$s_k^{(j,U_j)} = \begin{cases} J_k s_{k-1}^{(j,U_j)}, & j < k, \\ \alpha \partial_{W_k} f_{W_k}(y_{k-1})[\Delta U_k], & j = k, \\ 0, & j > k. \end{cases}$$

Since $s_0^{(j,U_j)} = 0$, we have $s_j^{(j,U_j)} = \alpha \partial_{W_j} f_{W_j}(y_{j-1})[\Delta U_j]$, and iterating,

$$s_K^{(j,U_j)} = J_K J_{K-1} \cdots J_{j+1}\, \alpha \partial_{W_j} f_{W_j}(y_{j-1})[\Delta U_j].$$

By Corollary 4 (see equation 3),

$$\left\| \partial_{W_j} f_{W_j}(y_{j-1})[\Delta U_j] \right\|_H \le C_W \left\| \Delta U_j \right\|_P \le C_W.$$

By Corollary 3,

$$\left\| s_K^{(j,U_j)} \right\|_H \le C_J \alpha C_W = \frac{C_s}{K}.$$

Summing over j,

$$\left\| \sum_{j=1}^{K} s_K^{(j,U_j)} \right\|_H \le \sum_{j=1}^{K} \left\| s_K^{(j,U_j)} \right\|_H \le K \cdot \frac{C_s}{K} = C_s.$$

□

Corollary 5 (Uniform Bound on First-Order Directional Derivatives)**.** From the proof above, it follows immediately that for all $1 \le j \le k \le K$,

$$\left\| s_k^{(j,U_j)} \right\|_H \le \frac{C_s}{K}, \qquad C_s := C_J \beta C_W.$$

In particular, for $k = K$,

$$\left\| s_K^{(j,U_j)} \right\|_H \le \frac{C_s}{K}, \qquad \left\| \sum_{j=1}^{K} s_K^{(j,U_j)} \right\|_H \le C_s.$$

B.3 Second-Order Directional Derivative Bound

For all $1 \le j, \ell \le K$,

$$\left\| h_K^{(j,\ell;U_j,V_\ell)} \right\|_H \le \begin{cases} C_{h1}/K^2, & j \ne \ell, \\ C_{h2}/K, & j = \ell, \end{cases}$$

with $C_{h1}, C_{h2} = \Theta(1)$. Consequently,

$$\left\| \sum_{j=1}^{K} \sum_{\ell=1}^{K} h_K^{(j,\ell;U_j,V_\ell)} \right\|_H \le C_h. \tag{6}$$

Proof. We proceed in three stages: derive the recursion, expand to layer K, and estimate.

Stage 1: Recursion. Differentiating the recursion for $s_k^{(j,U_j)}$ with respect to W_ℓ in direction ΔV_ℓ,

$$\begin{aligned} h_k^{(j,\ell;U_j,V_\ell)} = {} & J_k h_{k-1}^{(j,\ell;U_j,V_\ell)} + \alpha \partial_y^2 f_{W_k}(y_{k-1})[s_{k-1}^{(j,U_j)}, s_{k-1}^{(\ell,V_\ell)}] \\ & + \alpha \mathbf{1}_{j=k} \partial_y \partial_{W_k} f_{W_k}(y_{k-1})[s_{k-1}^{(\ell,V_\ell)}, \Delta U_k] \\ & + \alpha \mathbf{1}_{\ell=k} \partial_y \partial_{W_k} f_{W_k}(y_{k-1})[s_{k-1}^{(j,U_j)}, \Delta V_k] \\ & + \alpha \mathbf{1}_{j=k} \mathbf{1}_{\ell=k} \partial_{W_k}^2 f_{W_k}(y_{k-1})[\Delta U_k, \Delta V_k]. \end{aligned}$$

Since $h_0^{(j,\ell;U_j,V_\ell)} = 0$ and $s_{k-1}^{(j,U_j)} = 0$ for $j \ge k$, we have:

- Case A ($j, \ell < k$): $h_k = J_k h_{k-1} + \alpha \partial_y^2 f[s_{k-1}^{(j)}, s_{k-1}^{(\ell)}]$.
- Case B ($j = k, \ell < k$): $h_k = \alpha \partial_y \partial_{W_k} f[s_{k-1}^{(\ell)}, \Delta U_k]$.
- Case C ($\ell = k, j < k$): $h_k = \alpha \partial_y \partial_{W_k} f[s_{k-1}^{(j)}, \Delta V_k]$.
- Case D ($j = \ell = k$): $h_k = \alpha \partial_{W_k}^2 f[\Delta U_k, \Delta V_k]$.
- Case E ($j > k$ or $\ell > k$): $h_k = 0$.

Stage 2: Closed-form expansion. We unroll the recursion from the first nonzero layer to layer K.

Case $j < \ell$. The first nonzero term occurs at $k = \ell$ in Case C:

$$h_\ell^{(j,\ell;U_j,V_\ell)} = \alpha \partial_y \partial_{W_\ell} f_{W_\ell}(y_{\ell-1})[s_{\ell-1}^{(j,U_j)}, \Delta V_\ell].$$

For $k = \ell + 1, \dots, K$, Case A adds the source term $\alpha \partial_y^2 f_{W_k}[s_{k-1}^{(j,U_j)}, s_{k-1}^{(\ell,V_\ell)}]$.

Iterating,

$$\begin{aligned} h_K^{(j,\ell;U_j,V_\ell)} &= J_K J_{K-1} \cdots J_{\ell+1}\, \alpha \partial_y \partial_{W_\ell} f_{W_\ell}(y_{\ell-1})[s_{\ell-1}^{(j,U_j)}, \Delta V_\ell] \\ &\quad + \sum_{k=\ell+1}^{K} J_K J_{K-1} \cdots J_{k+1}\, \alpha \partial_y^2 f_{W_k}(y_{k-1})[s_{k-1}^{(j,U_j)}, s_{k-1}^{(\ell,V_\ell)}]. \end{aligned}$$

Case $j > \ell$. By symmetry, $h_K^{(j,\ell;U_j,V_\ell)} = h_K^{(\ell,j;V_\ell,U_j)}$.

Case $j = \ell > 0$. The first nonzero term occurs at $k = j$ in Case D:

$$h_j^{(j,j;U_j,V_j)} = \alpha \partial_{W_j}^2 f_{W_j}(y_{j-1})[\Delta U_j, \Delta V_j].$$

Iterating,

$$\begin{aligned} h_K^{(j,j;U_j,V_j)} &= J_K J_{K-1} \cdots J_{j+1}\, \alpha \partial_{W_j}^2 f_{W_j}(y_{j-1})[\Delta U_j, \Delta V_j] \\ &\quad + \sum_{k=j+1}^{K} J_K J_{K-1} \cdots J_{k+1}\, \alpha \partial_y^2 f_{W_k}(y_{k-1})[s_{k-1}^{(j,U_j)}, s_{k-1}^{(j,V_j)}]. \end{aligned}$$

Stage 3: Estimates. By Corollary 4 and the bound on $s_{k-1}^{(j,U_j)}$ in Corollary 5, we have:

$$\begin{aligned} \left\| \alpha \partial_y^2 f_{W_k}[s_{k-1}^{(j,U_j)}, s_{k-1}^{(\ell,V_\ell)}] \right\|_H &\le \alpha M_f \frac{C_s^2}{K^2} = \frac{C_{yy}}{K^3}, \\ \left\| \alpha \partial_y \partial_{W_\ell} f_{W_\ell}[s_{\ell-1}^{(j,U_j)}, \Delta V_\ell] \right\|_H &\le \alpha M_{Wf} \frac{C_s}{K} = \frac{C_{yW}}{K^2}, \\ \left\| \alpha \partial_{W_j}^2 f_{W_j}[\Delta U_j, \Delta V_j] \right\|_H &\le \alpha M_{WW} = \frac{C_{WW}}{K}. \end{aligned}$$

By Corollary 3, $\|J_K \cdots J_{k+1}\|_{H \to H} \le C_J$.

Case $j \neq \ell$. Using the expansion for $j < \ell$ (the case $j > \ell$ is symmetric),

$$\left\| h_K^{(j,\ell;U_j,V_\ell)} \right\|_H \le C_J \frac{C_{yW}}{K^2} + \sum_{k=\ell+1}^{K} C_J \frac{C_{yy}}{K^3} \le \frac{C_J C_{yW}}{K^2} + K \frac{C_J C_{yy}}{K^3} = \frac{C_J (C_{yW} + C_{yy})}{K^2} =: \frac{C_{h1}}{K^2}.$$

Case $j = \ell$. For $j > 0$,

$$\left\| h_K^{(j,j;U_j,V_j)} \right\|_H \le C_J \frac{C_{WW}}{K} + \sum_{k=j+1}^{K} C_J \frac{C_{yy}}{K^3} \le \frac{C_J C_{WW}}{K} + \frac{C_J C_{yy}}{K^2} \le \frac{C_J (C_{WW} + C_{yy})}{K} =: \frac{C_{h2}}{K}.$$

There are $K(K-1)$ pairs with $j \neq \ell$ and K pairs with $j = \ell$:

$$\left\| \sum_{j=1}^{K} \sum_{\ell=1}^{K} h_K^{(j,\ell;U_j,V_\ell)} \right\|_H \le K(K-1) \frac{C_{h1}}{K^2} + K \frac{C_{h2}}{K} \le C_{h1} + C_{h2} =: C_h.$$

□

B.4 LEMMAS

Lemma 1 (Directional Derivative Lipschitz Bound)**.** Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ be differentiable and $\|\cdot\|_B$ be any matrix norm. If for all Z on the segment between Z^1, Z^2 and all $\|\Delta U\|_B \leq 1$, $|\nabla f(Z)[\Delta U]| \leq L$, then

$$|f(Z^1) - f(Z^2)| \leq L\|Z^1 - Z^2\|_B.$$

Proof. If $Z^1 = Z^2$, the claim is trivial. Assume $Z^1 \neq Z^2$. Let $Z(t) = Z^2 + (Z^1 - Z^2)t$ for $t \in [0, 1]$. By the fundamental theorem of calculus,

$$f(Z^1) - f(Z^2) = \int_0^1 \nabla f(Z(t))[Z^1 - Z^2]\, dt.$$

Set $\Delta U = (Z^1 - Z^2)/\|Z^1 - Z^2\|_B$. Then $\|\Delta U\|_B = 1$, and

$$|\nabla f(Z(t))[Z^1 - Z^2]| = \|Z^1 - Z^2\|_B\, |\nabla f(Z(t))[\Delta U]| \leq L\|Z^1 - Z^2\|_B.$$

Integrating yields the result. □

Lemma 2 (Directional Hessian Smoothness Bound)**.** Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ be twice differentiable and $\|\cdot\|_B$ be any matrix norm. If for all Z on the segment between Z^1, Z^2 and all $\|\Delta U\|_B, \|\Delta V\|_B \leq 1$, $|\nabla^2 f(Z)[\Delta U, \Delta V]| \leq L$, then

$$\|\nabla f(Z^1) - \nabla f(Z^2)\|_{B,*} \leq L\|Z^1 - Z^2\|_B.$$

Proof. If $Z^1 = Z^2$, the claim is trivial. Assume $Z^1 \neq Z^2$. For any direction ΔU,

$$\nabla f(Z^1)[\Delta U] - \nabla f(Z^2)[\Delta U] = \int_0^1 \nabla^2 f(Z(t))[\Delta U, Z^1 - Z^2]\, dt.$$

By the definition of the dual norm,

$$\|\nabla f(Z^1) - \nabla f(Z^2)\|_{B,*} = \sup_{\|\Delta U\|_B \leq 1} |\nabla f(Z^1)[\Delta U] - \nabla f(Z^2)[\Delta U]|.$$

Set $\Delta V = (Z^1 - Z^2)/\|Z^1 - Z^2\|_B$. Then $\|\Delta V\|_B = 1$, and

$$|\nabla^2 f(Z(t))[\Delta U, Z^1 - Z^2]| = \|Z^1 - Z^2\|_B\, |\nabla^2 f(Z(t))[\Delta U, \Delta V]| \leq L\|Z^1 - Z^2\|_B.$$

Integrating and taking the supremum yields the result. □

B.5 PROOF OF THE THEOREM

Theorem (Depth-Independent Lipschitz Continuity)**.** Under Assumptions 1 to 3 and $\alpha = \beta/K$, $\beta = \Theta(1)$, there exists $L_{\text{Lip}} = O(1)$, independent of K, such that

$$\sup_{\|\Delta U\|_{\text{block}} \leq 1} |\nabla \mathcal{L}(W)[\Delta U]| \leq L_{\text{Lip}}.$$

Consequently,

$$|\mathcal{L}(W^1) - \mathcal{L}(W^2)| \leq L_{\text{Lip}}\|W^1 - W^2\|_{\text{block}}.$$

Proof. By the chain rule,

$$\nabla_W \mathcal{L}(W)[\Delta U] = l'(y_K) \sum_{j=1}^{K} s_K^{(j,U_j)}.$$

By Assumption 3 and Corollary 5,

$$|\nabla_W \mathcal{L}(W)[\Delta U]| \leq L_J C_s =: L_{\text{Lip}}.$$

Lemma 1 gives the Lipschitz consequence. □

Theorem (Depth-Independent L-Smoothness)**.** Under Assumptions 1 to 3 and $\alpha = \beta/K$, $\beta = \Theta(1)$, there exists $L = O(1)$, independent of K, such that

$$\sup_{\|\Delta U\|_{\text{block}}\le 1, \|\Delta V\|_{\text{block}}\le 1} |\nabla^2 \mathcal{L}(W)[\Delta U, \Delta V]| \le L.$$

Consequently,

$$\|\nabla \mathcal{L}(W^1) - \nabla \mathcal{L}(W^2)\|_{\text{block},*} \le L \|W^1 - W^2\|_{\text{block}}.$$

Proof. By the chain rule,

$$\nabla^2_W \mathcal{L}[\Delta U, \Delta V] = l''(y_K)\Big(\sum_{j=1}^K s_K^{(j,U_j)}\Big)\Big(\sum_{\ell=1}^K s_K^{(\ell,V_\ell)}\Big) + l'(y_K)\sum_{j=1}^K\sum_{\ell=1}^K h_K^{(j,\ell;U_j,V_\ell)}.$$

By Assumption 3, Corollary 5, and equation 6,

$$|\nabla^2_W \mathcal{L}[\Delta U, \Delta V]| \le M_J C_s^2 + L_J C_h =: L.$$

Lemma 2 gives the smoothness consequence. □

C Discussion

In this section we verify that the hypotheses of Theorem 1 and Theorem 2 — in particular, the residual-function regularity of Assumption 2 — are met by the building blocks used in practice. We first treat multi-layer perceptron (MLP) residual blocks and record exactly which part of Assumption 2 needs to hold only locally; we also show that attention blocks satisfy the same regularity at fixed sequence length and feature dimensions. Throughout, $H = \mathbb{R}^d$ is equipped with the (p, mean) norm $\|x\|_H = d^{-1/p}\|x\|_p$ for some $1 \le p \le \infty$, and $\|\cdot\|_{H\to H}$ denotes the $(p,\text{mean}) \to (p,\text{mean})$ operator norm. We use $\|\cdot\|_{H\to H}$ for the induced operator norm and for the parameter space P we measure matrix components with $\|\cdot\|_{(p,\text{mean})\to(q,\text{mean})}$ for the prescribed matrix norm from (p,mean) to (q,mean) where $p \le q$. By mean normalization, we have

$$\|A\|_{H\to H} \le \|A\|_{(p,\text{mean})\to(q,\text{mean})}, \qquad A \in \mathbb{R}^{d\times d}.$$

Thus, bounds in the prescribed matrix norm also control hidden-state propagation. Moreover,

$$\|\operatorname{Diag}(v)\|_{H\to H} = \|v\|_\infty, \qquad \|\mathbf{1}\|_H = 1,$$

where $\mathbf{1} = (1,\dots,1)^\top$. The hidden dimension d and the norm exponents p, q are fixed independently of the residual depth K.

C.1 Multi-Layer Neural Networks

Remark 2 (Localization of the parameter-regularity conditions)**.** In the proofs of Theorem 1 and Theorem 2, the W-regularity conditions of Assumption 2 — namely the bounds equation 3–equation 5 on $\partial_{W_k} f_{W_k}$, $\partial_y \partial_{W_k} f_{W_k}$, and $\partial^2_{W_k} f_{W_k}$ — are invoked only at the iterates y_{k-1}, $k = 1,\dots,K$. By Corollary 1, every iterate satisfies $\|y_{k-1}\|_H \le C_y$. Consequently, it suffices for the theorems that these three conditions hold only for y in the ball $\{\|y\|_H \le C_y\}$, rather than for all $y \in H$. The y-regularity conditions (the first two lines of Assumption 2, plus the bound $\|f_{W_k}(y_0)\|_H \le C_f$) are genuinely global and are unaffected.

Proposition 1 (MLP residual satisfies Assumption 2)**.** Let $H = \mathbb{R}^d$ and let the residual function be the N-layer MLP

$$f_{W_k}(y) = \sigma\big(W^{(N)}\sigma(\cdots\sigma(W^{(1)}y + b^{(1)})\cdots) + b^{(N)}\big),$$

where $W_k = (W^{(1)},\dots,W^{(N)}, b^{(1)},\dots,b^{(N)})$ with $W^{(1)},\dots,W^{(N)} \in \mathbb{R}^{d\times d}$ and $b^{(1)},\dots,b^{(N)} \in \mathbb{R}^d$, and equip P with the norm

$$\|W_k\|_P = \max\Big\{\|W^{(1)}\|_{(p,\text{mean})\to(q,\text{mean})},\dots,\|W^{(N)}\|_{(p,\text{mean})\to(q,\text{mean})}, \Big\|b^{(1)}\Big\|_H,\dots,\Big\|b^{(N)}\Big\|_H\Big\}.$$

Suppose that

1. σ is twice continuously differentiable, with constants chosen so that
$$\max\{1, |\sigma(0)|, \sup_{t\in\mathbb{R}} |\sigma'(t)|\} \le L_\sigma, \qquad d^{1/p} \sup_{t\in\mathbb{R}} |\sigma''(t)| \le M_\sigma,$$
where $d^{1/\infty} = 1$;
2. the internal depth N is fixed and does not grow with K;
3. for all $\ell = 1, \dots, N$, $\|W^{(\ell)}\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})} \le B_W$ and $\left\|b^{(\ell)}\right\|_H \le B_b$, with B_W, B_b independent of K.

Without loss of generality, enlarge the weight bound so that $B_W \ge 1$. The admissible parameter domain is the convex set
$$\Omega = \left\{ W_k \in P : \|W^{(\ell)}\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})} \le B_W,\ \left\|b^{(\ell)}\right\|_H \le B_b, \quad \ell = 1, \dots, N \right\}.$$
All parameter-regularity comparisons are restricted to Ω.

Then f_{W_k} satisfies Assumption 2 in the localized sense of Remark 2: the y-regularity conditions hold globally, while the W-regularity conditions hold on the ball $\{\|y\|_H \le C_y\}$, all with constants $L_f, M_f, C_W, M_{Wf}, M_{WW}, C_f$ depending only on $N, L_\sigma, M_\sigma, B_W, B_b, C_x, C_y$, and not on K.

Proof. Forward pass. Introduce the pre-activations $z^{(\ell)}$ and activations $a^{(\ell)}$ by
$$a^{(0)} = y, \qquad z^{(\ell)} = W^{(\ell)} a^{(\ell-1)} + b^{(\ell)}, \qquad a^{(\ell)} = \sigma(z^{(\ell)}), \qquad f_{W_k}(y) = a^{(N)}.$$
By mean-normalization we have that $\|W\|_{H\to H} \le \|W\|_{(p,mean)\to(q,mean)} \le B_W$.Write $D^{(\ell)} := \mathrm{Diag}(\sigma'(z^{(\ell)}))$ and $G^{(\ell)} := D^{(\ell)} W^{(\ell)}$ for the per-layer Jacobian, so that
$$\left\|D^{(\ell)}\right\|_{H\to H} \le L_\sigma, \qquad \left\|G^{(\ell)}\right\|_{H\to H} \le L_\sigma B_W$$
hold globally (for all y). Let $F^{(\ell)} := G^{(N)} G^{(N-1)} \cdots G^{(\ell+1)}$ denote forward propagation from layer ℓ to the output (with $F^{(N)} = I$), so that
$$\left\|F^{(\ell)}\right\|_{H\to H} \le (L_\sigma B_W)^{N-\ell}.$$

Bounded activations. By the mean value theorem applied componentwise to $a^{(\ell)} = \sigma(z^{(\ell)})$, there exists a diagonal matrix $\tilde{D}^{(\ell)} = \mathrm{Diag}(\sigma'(\xi_1), \dots, \sigma'(\xi_d))$ with $\left\|\tilde{D}^{(\ell)}\right\|_{H\to H} \le L_\sigma$ such that $a^{(\ell)} = \sigma(0)\mathbf{1} + \tilde{D}^{(\ell)} z^{(\ell)}$. Since $\|\mathbf{1}\|_H = 1$, the triangle inequality gives
$$\left\|a^{(\ell)}\right\|_H \le |\sigma(0)| + L_\sigma \left\|z^{(\ell)}\right\|_H.$$
Combining this with $\left\|z^{(\ell)}\right\|_H \le B_W \left\|a^{(\ell-1)}\right\|_H + B_b$ and $\left\|a^{(0)}\right\|_H = \|y\|_H \le C_y$, an induction over the $N = O(1)$ layers yields a constant $R = R(N, L_\sigma, B_W, B_b, C_y) \ge 1$ such that
$$\left\|a^{(\ell)}\right\|_H \le R, \qquad \ell = 0, \dots, N,$$
$$\left\|z^{(\ell)}\right\|_H \le R, \qquad \ell = 1, \dots, N, \qquad \text{on } \{\|y\|_H \le C_y\}.$$

(i) Global y-regularity. The chain rule gives
$$\partial_y f_{W_k}(y) = G^{(N)} G^{(N-1)} \cdots G^{(1)},$$
hence
$$\|\partial_y f_{W_k}(y)\|_{H\to H} \le (L_\sigma B_W)^N =: L_f \qquad \text{for all } y \in H,$$
which is equation 1; the L_f-Lipschitz condition in y follows by the mean value theorem.

For the second derivative bound, expand the difference of the Jacobian products as a telescoping sum: for $y_1, y_2 \in H$,

$$\partial_y f(y_1) - \partial_y f(y_2) = \sum_{\ell=1}^{N} G^{(N)}(y_1) \cdots G^{(\ell+1)}(y_1) \left[G^{(\ell)}(y_1) - G^{(\ell)}(y_2)\right] G^{(\ell-1)}(y_2) \cdots G^{(1)}(y_2).$$

Since $G^{(\ell)} = D^{(\ell)} W^{(\ell)}$ and $\partial_y z^{(\ell)} = W^{(\ell)} G^{(\ell-1)} \cdots G^{(1)}$, each factor is Lipschitz in y:

$$\begin{aligned}\left\|G^{(\ell)}(y_1) - G^{(\ell)}(y_2)\right\|_{H\to H} &\le \left\|D^{(\ell)}(y_1) - D^{(\ell)}(y_2)\right\|_{H\to H} \left\|W^{(\ell)}\right\|_{H\to H} \\ &\le M_\sigma \left\|z^{(\ell)}(y_1) - z^{(\ell)}(y_2)\right\|_H B_W \\ &\le M_\sigma B_W^2 (L_\sigma B_W)^{\ell-1} \|y_1 - y_2\|_H .\end{aligned}$$

Substituting into the Leibniz sum (each term carries $N-1$ remaining factors of norm at most $L_\sigma B_W$) yields $\|\partial_y f(y_1) - \partial_y f(y_2)\|_{H\to H} \le M_f \|y_1 - y_2\|_H$ with

$$M_f := M_\sigma B_W^2 \sum_{\ell=1}^{N} (L_\sigma B_W)^{N+\ell-2},$$

a constant independent of K; equivalently equation 2 holds. This verifies the two global y-conditions of Assumption 2.

(ii) Localized W-regularity on the ball $\{\|y\|_H \le C_y\}$. Differentiating through layer ℓ, the derivative with respect to the weight and bias blocks is

$$\begin{aligned}\partial_{W^{(\ell)}} f_{W_k}(y)[\Delta W^{(\ell)}] &= F^{(\ell)} D^{(\ell)} \Delta W^{(\ell)} a^{(\ell-1)}, \\ \partial_{b^{(\ell)}} f_{W_k}(y)[\Delta b^{(\ell)}] &= F^{(\ell)} D^{(\ell)} \Delta b^{(\ell)}.\end{aligned}$$

Using $\left\|F^{(\ell)}\right\|_{H\to H} \le (L_\sigma B_W)^{N-\ell}$, $\left\|D^{(\ell)}\right\|_{H\to H} \le L_\sigma$, $\left\|a^{(\ell-1)}\right\|_H \le R$, and $\|\Delta W_k\|_P \le 1$ (so each block has norm at most 1), we get

$$\|\partial_{W_k} f_{W_k}(y)\|_{P\to H} \le \sum_{\ell=1}^{N} (L_\sigma B_W)^{N-\ell} L_\sigma (R+1) =: C_W,$$

which is equation 3; the mean value theorem over the convex parameter set Ω gives the C_W-Lipschitz condition in W.

For equation 4, differentiate $\partial_y f = G^{(N)} \cdots G^{(1)}$ with respect to $W^{(\ell)}$ (for $m \ge \ell$, $G^{(m)}$ depends on $W^{(\ell)}$ through $z^{(m)}$; $G^{(\ell)}$ depends additionally on $W^{(\ell)}$ directly):

$$\partial_{W^{(\ell)}} \partial_y f\,[\Delta W^{(\ell)}] = \sum_{m=\ell}^{N} G^{(N)} \cdots G^{(m+1)}\, \partial_{W^{(\ell)}} G^{(m)}[\Delta W^{(\ell)}]\, G^{(m-1)} \cdots G^{(1)}.$$

We bound the differentiated factor on the ball, taking $\|\Delta W^{(\ell)}\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})} \le 1$. For $m = \ell$, differentiating $G^{(\ell)} = D^{(\ell)} W^{(\ell)}$ gives

$$\begin{aligned}\left\|\partial_{W^{(\ell)}} G^{(\ell)}[\Delta W^{(\ell)}]\right\|_{H\to H} &\le \left\|\partial_{W^{(\ell)}} D^{(\ell)}[\Delta W^{(\ell)}]\right\|_{H\to H} \left\|W^{(\ell)}\right\|_{H\to H} + \left\|D^{(\ell)}\right\|_{H\to H} \left\|\Delta W^{(\ell)}\right\|_{H\to H} \\ &\le M_\sigma R\, B_W + L_\sigma,\end{aligned}$$

where the two terms come from differentiating $D^{(\ell)}$ and $W^{(\ell)}$, respectively. For $m > \ell$, only the $D^{(m)}$-factor depends on $W^{(\ell)}$, through

$$\partial_{W^{(\ell)}} z^{(m)}[\Delta W^{(\ell)}] = W^{(m)} G^{(m-1)} \cdots G^{(\ell+1)} D^{(\ell)} \Delta W^{(\ell)} a^{(\ell-1)},$$

with the intermediate product interpreted as the identity when $m = \ell + 1$, so that

$$\begin{aligned}\left\|\partial_{W^{(\ell)}} G^{(m)}[\Delta W^{(\ell)}]\right\|_{H\to H} &\le M_\sigma \left\|\partial_{W^{(\ell)}} z^{(m)}[\Delta W^{(\ell)}]\right\|_H \left\|W^{(m)}\right\|_{H\to H} \\ &\le M_\sigma B_W (L_\sigma B_W)^{m-\ell} R.\end{aligned}$$

Setting

$$\Gamma := \max\{L_\sigma B_W,\ L_\sigma + M_\sigma R B_W,\ M_\sigma B_W (L_\sigma B_W)^{N-1} R\},$$

each summand above is a product of N factors of norm at most Γ. The same bounds apply to unit bias directions, replacing $\Delta W^{(\ell)} a^{(\ell-1)}$ by $\Delta b^{(\ell)}$ and omitting the direct weight-derivative term, since $R \geq 1$. Summing the at most N summands and accounting for the N weight and N bias blocks gives

$$\|\partial_y \partial_{W_k} f_{W_k}(y)\|_{H \times P \to H} \leq 2N^2 \Gamma^N =: M_{Wf};$$

the mean value theorem on the hidden-state ball yields equation 4.

(iii) Second-order W-regularity. The map $\partial_{W_k} f_{W_k}(y)$ is the sum over $\ell = 1, \ldots, N$ of the two blocks displayed in *(ii)*. Differentiating each block once more with respect to W_k applies the product rule a second time, so every resulting term is a product of factors, each of which is one of $\sigma'(z^{(m)})$, $\sigma''(z^{(m)})$, a weight matrix $W^{(m)}$, an activation $a^{(m)}$, the input y, or a direction $\Delta U_k, \Delta V_k$. Take $\|\Delta U_k\|_P, \|\Delta V_k\|_P \leq 1$. By the choice of M_σ, the second-order activation terms satisfy

$$\|\sigma''(z) \odot u \odot v\|_H \leq d^{1/p} \sup_{t \in \mathbb{R}} |\sigma''(t)|\ \|u\|_H\ \|v\|_H \leq M_\sigma\ \|u\|_H\ \|v\|_H,$$

where $\odot$ denotes componentwise multiplication. On the ball $\{\|y\|_H \leq C_y\}$ every such factor is bounded, and the total number of terms is bounded by a constant depending only on N. Hence

$$\left\|\partial^2_{W_k} f_{W_k}(y)\right\|_{P \times P \to H} \leq M_{WW} \qquad \text{on } \{\|y\|_H \leq C_y\},$$

and the mean value theorem over Ω gives equation 5. All constants depend only on $N, L_\sigma, M_\sigma, B_W, B_b, C_y$.

(iv) The initial bound. Since $\|y_0\|_H = \|x\|_H \leq C_x$, rerunning the bounded-activation induction on the ball $\{\|y\|_H \leq C_x\}$ (in place of C_y) gives

$$\|f_{W_k}(y_0)\|_H = \left\|a^{(N)}\right\|_H \leq R_x =: C_f, \qquad R_x = R(N, L_\sigma, B_W, B_b, C_x),$$

which does not depend on C_y. This completes the verification of Assumption 2 in the localized sense of Remark 2. □

Corollary 6 (MLP residual networks satisfy the main theorems)**.** Let the residual function of Definition 1 be the MLP of Proposition 1, and suppose Assumption 1 and Assumption 3 hold (bounded input x; bounded, smooth loss l). Then the loss $\mathcal{L}$ satisfies the conclusions of Theorem 1 and Theorem 2: there exist constants $L_{\mathrm{Lip}}, L = O(1)$, independent of the depth K, such that $\mathcal{L}$ is L_{Lip}-Lipschitz and L-smooth in the block norm.

Remark 3 (Scope of the remaining assumptions)**.** Assumption 1 is a data-normalization hypothesis (bounded inputs and unit perturbation directions, the latter by homogeneity). Assumption 3 is a hypothesis on the loss function l only, not on the network, and is satisfied by standard choices (e.g. logistic or quadratic losses on a bounded domain). Thus the only network-dependent hypothesis — Assumption 2 — is exactly what Proposition 1 establishes for MLP residual blocks.

C.2 Attention Blocks

Proposition 1 verifies Assumption 2 for MLP residual blocks equipped with the $p \to q$ operator norm on $H = \mathbb{R}^d$. A natural question is whether Assumption 2 — and therefore the depth-independent guarantees of Theorems 1 and 2 — also holds for attention blocks. (Xu et al., 2026) answer this in the affirmative from the matrix-operator-norm perspective, which we summarize here.

A single attention head with query, key, value, and output projections W_q, W_k, W_v, W_o maps a sequence of token features $X \in \mathbb{R}^{s \times d}$ to

$$\mathrm{Att}(X) = \mathrm{softmax}\Big(\frac{X W_q W_k^\top X^\top}{\sqrt{d_v}}\Big) X W_v W_o,$$

where s is the sequence length and d_v the per-head value dimension. The regularity of this map is governed by the scale of the attention logits $X W_q W_k^\top X^\top$: to keep the softmax L-smooth, this scale

must not grow with the width. (Xu et al., 2026) control it by row or column normalization of the weight matrices.

Row normalization (geometry $(p, \mathrm{mean}) \to \ell_\infty$). The query and key vectors are bounded in ℓ_∞, so each query–key inner product satisfies

$$|q^\top k| \le d_v \, \|q\|_\infty \|k\|_\infty = O(d_v),$$

and scaling the logits by $1/d_v$ keeps them $O(1)$ in the width.

Column normalization (geometry $(1, \mathrm{mean}) \to (q, \mathrm{mean})$). The inner product scales as $q^\top k = O\big(d_v^{\max\{1,2/q\}}\big)$, so scaling the logits by $d_v^{-\max\{1,2/q\}}$ keeps them width-independent.

With the logits thus controlled, the loss of a K-layer network composed of such blocks is Lipschitz and smooth with constants independent of the hidden width, under bounded inputs and parameters and bounded first/second derivatives of the activation and loss. This is precisely the residual-function regularity demanded by Assumption 2, expressed in the mean-normalized operator-norm geometry. We record this as follows.

Proposition 2 (Attention blocks satisfy the residual regularity)**.** Under row or column normalization of W_q, W_k, W_v, W_o and bounded activations, the attention block satisfies Assumption 2

Proof sketch. Row normalization bounds the query and key vectors in ℓ_∞, so the logits are $O(d_v)$ and the $1/d_v$ scaling makes the argument of the softmax $O(1)$. This scaling controls the magnitude of the logits. To establish regularity of the full attention branch, we also account for its dependence on the hidden states and projection parameters. Column normalization bounds the query and key vectors in the (q, mean) norm, giving logits of size $O(d_v^{\max\{1,2/q\}})$, which is compensated by the $d_v^{-\max\{1,2/q\}}$ scaling. For fixed sequence length and feature dimensions, the attention branch is jointly twice continuously differentiable in its input and projection parameters. On a bounded closed input domain and a bounded closed parameter domain, its first and second derivatives are therefore uniformly bounded in the prescribed norms. In particular, this bounds the state, parameter, and mixed derivatives required by Assumption 2. If the architecture and domain bounds are independent of the residual depth K, the resulting regularity constants are independent of K. On convex domains, the corresponding Lipschitz inequalities follow by integration along line segments. □

Thus, the MLP and attention branches satisfy the required regularity conditions on their respective admissible domains. The depth-independent conclusions of Theorems 1 and 2 apply provided the regularity bounds hold uniformly on a domain containing the hidden-state trajectories for all admissible parameter tuples, with parameter comparisons restricted to segments within the admissible parameter domain.

D ADDITIONAL EXPERIMENTS

Here we provide the validation loss curves of GPT-2 and the validation loss curves of ResNet.

D.1 GPT2 VAL LOSS

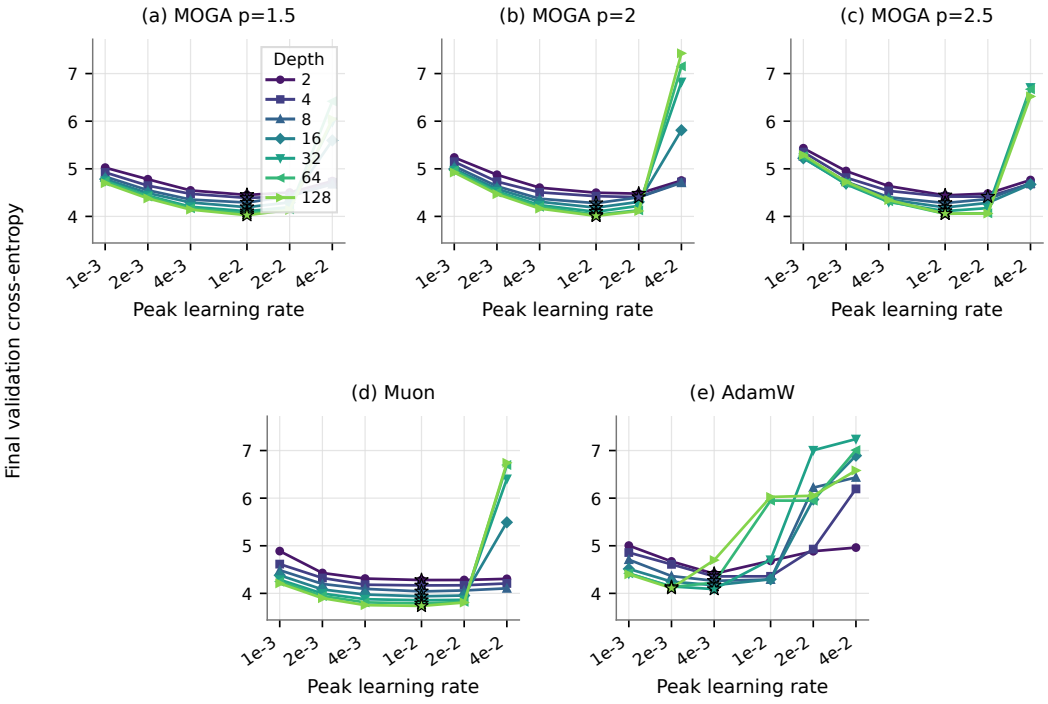

Figure 4: Learning-rate transfer across GPT-2 depths at fixed width. Panels show MOGA with $p \in \{1.5, 2, 2.5\}$, Muon, and AdamW. Colors and markers indicate transformer depth. Stars mark the lowest validation loss among the tested learning rates at each depth.

As shown in Figure 4, models that vary drastically in depth obtain their best validation loss performance at around the same peak learning rate. Muon and MOGA show strong learning rate transfer as all depths select nearly the same learning rate. The 4 Muon/MOGA panels show a very stable low loss region and a learning rate at a smaller depth can generally be transferred to a deeper model. AdamW exhibits the same instability for deeper models at the optimal learning rate $\eta^* = 4 \times 10^{-3}$ of the shallow models and the optimal learning rate shifts slightly towards the left. The optimal learning rates across 7 depths still stay within a tight region.

D.2 RESNET VAL-LOSS

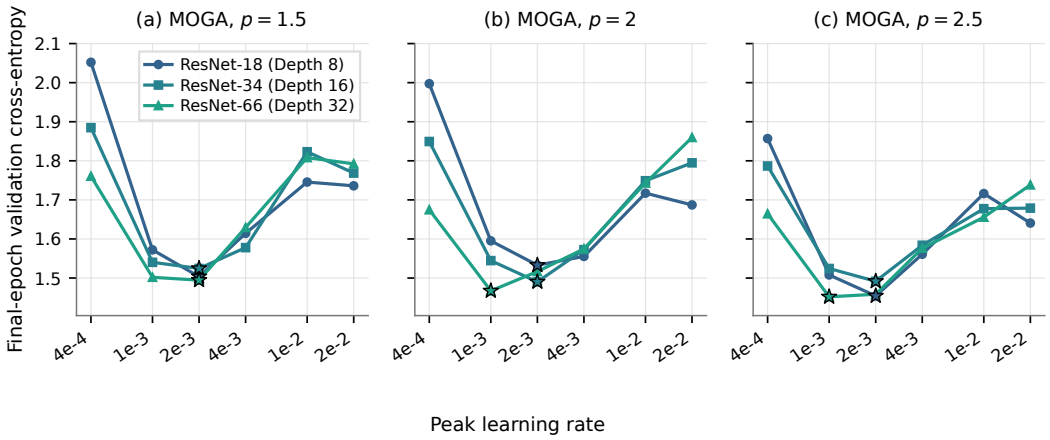

Figure 5: Learning-rate transfer across ResNet depths. Final validation loss is plotted against the peak learning rate, with each curve corresponding to a different depth. Colors and markers indicate ResNet depth. Stars mark the lowest validation loss among the tested learning rates at each depth.

The ResNet Validation loss curves shown in figure 5 show that the optimal learning rates shifts slightly towards the left for ResNet-66 ($K = 32$). MOGA $p = 1.5$ and MOGA $p = 2.5$ show nearly identical validation loss for $\eta = 1 \times 10^{-3}$ and $\eta = 2 \times 10^{-3}$, therefore the regret of selecting the optimal learning rate from shallower ResNets and transferring to ResNet-66 is low. However, MOGA $p = 2$ clearly shows a preference for the smaller learning rate $\eta = 1 \times 10^{-3}$. Furthermore, increasing depth does not consistently improve validation loss at a fixed learning rate.

The above shows that the optimal learning rate still stays in a very tight region, but learning rate transfer is not as stable as evaluated on training loss. The models exhibited overfitting during training, with improvements in training loss not consistently translating into improvements in validation loss. Accordingly, learning-rate transfer measured by training loss should be distinguished from transfer measured by validation loss.