CONVERGENCE GUARANTEES OF GRADIENT DESCENT FOR NEURAL NETWORKS VIA GENERALIZED LIPSCHITZ SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish the first convergence guarantees of gradient descent for general feed-forward neural networks of any width or depth, with any initialization or dataset. We only assume that the activation functions are linearly bounded, Lipschitz continuous, and Lipschitz smooth—properties that hold for linear, tanh, softplus, sigmoid, and smoothed ReLU functions—and that the loss function is Lipschitz smooth in the *model outputs*, a mild condition satisfied by mean squared error and binary cross-entropy loss. By relating the Lipschitz properties of one layer to the next, we obtain a novel generalized Lipschitz smoothness condition for an L-layer neural network where the change in gradient is upper bounded by the change in the parameter space, multiplied by a polynomial of the parameter norms at both endpoints of degree $2L-2$. This yields a descent lemma where the loss decreases as long as the learning rate is small enough with respect to the parameter norms. By ensuring that the leading term of the polynomial grows at a controlled rate, we prove that the minimum squared gradient norm converges to zero in T iterations at rate $O(1/T^{1/L})$.

1 INTRODUCTION

The behavior of gradient descent on the loss functions of neural networks has evaded a complete theoretical understanding for decades. These loss landscapes are not only highly nonconvex, but more critically, they do not satisfy *Lipschitz smoothness*, where the change in gradient is linearly bounded by the change in parameter space. Instead, the local Lipschitz constant, which dictates the sharpness of the landscape, can grow arbitrarily large as the parameters approach infinity. Because Lipschitz smoothness is a key condition for proving the convergence of gradient algorithms, a fundamental gap has emerged in machine learning research: popular optimization methods are analyzed under the smoothness assumption, yet empirically validated on neural network landscapes that are not theoretically well-characterized.

A rich body of research examines the convergence of gradient descent to *global minima* of extremely wide or overparametrized neural networks (Li & Liang, 2018; Du et al., 2019; Allen-Zhu et al., 2019; Xu et al., 2023; Taheri et al., 2025; Arora et al., 2019a; Chatterjee, 2026; Dana et al., 2025), with additional constraints on the architecture, activation functions, proportion of trained parameters, initialization scheme, or dataset. Under such conditions, the iterates remain within a bounded set close to the initialization, effectively satisfying *local* Lipschitz smoothness. While these analyses provide valuable insights into how gradient descent can achieve unexpectedly strong performance in deep learning, they do not explain the behavior of gradient descent beyond these strict assumptions. In particular, the overparametrized or infinite-width regime generally only captures the behavior of *lazy training*, where the model effectively behaves like a fixed kernel machine or linear model. In reality, finite-width neural networks undergo *feature learning*, where the weights may change significantly as the model learns feature representations of the data (Arora et al., 2019b; Yehudai & Shamir, 2019; Yang & Hu, 2022; Vyas et al., 2023; Chizat et al., 2019; Woodworth et al., 2020).

A separate line of research seeks to identify weaker requirements, such as *non-uniform* or *generalized Lipschitz* smoothness conditions, that still allow gradient descent to converge. For twice-differentiable functions, Lipschitz smoothness is equivalent to requiring that the Hessian norm is

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.

bounded by a constant. One relaxation is (L_0, L_1)-smoothness (Zhang et al., 2020), which requires that the Hessian norm is upper bounded by an affine function of the gradient norm. Other works consider bounding the Hessian with any nondecreasing function of the gradient norm (Li et al., 2023) or an affine function of the objective value (Vaswani et al., 2026). A related concept is *relative smoothness*, where the Hessian is upper bounded by the Hessian of some convex reference function, allowing for convergence of gradient descent or mirror descent (Bauschke et al., 2017; Lu et al., 2018). However, while these conditions can capture more complex settings, they are only *empirically* validated, and are not actually satisfied by the loss landscapes of deep neural networks.

In this work, we provide a fresh analytical framework addressing the limitations of the above two approaches. We explicitly derive a generalized Lipschitz smoothness condition, which we call *double polynomial smoothness*, that fully characterizes L-layer feedforward neural networks of any width, as long as the activation functions are linearly bounded, Lipschitz continuous, and Lipschitz smooth, and the loss is Lipschitz smooth in the model outputs[1]. By recursively relating the Lipschitz properties of one layer to the next, we establish that the change in model function gradient is upper bounded by the change in parameter space, multiplied by a polynomial of the parameter norms at both endpoints of degree $2L - 2$. This property can be further passed to the loss function, such that by integrating over the line segment between the endpoints, we obtain a novel descent lemma where the loss decreases as long as the learning rate is small enough with respect to the parameter norms and loss at each step. The last step is showing with induction that the leading term of the polynomial grows at most sublinearly to infinity, at rate $T^{\frac{L-1}{L}}$. This allows us to prove that the minimum squared gradient norm converges to zero at rate $O(1/T^{1/L})$.

We examine the practical implications of our theory by conducting experiments in training feedforward models on the CIFAR-10 image classification dataset. We observe that scaling the learning rate according to our theory leads to faster and more stable convergence compared to constant learning rates by correctly adapting to the local geometry. Moreover, we observe that increasing the number of layers indeed slows the convergence, suggesting that the derived dependence on L is fundamental and not just an artifact of the proof. Code is open-sourced at the anonymous Github repository at `https://anonymous.4open.science/r/nn_optimization-82F8/README.md`.

Our convergence analysis applies to a broad range of feedforward neural networks, including models with linear, tanh, softplus, sigmoid, and smoothed ReLU (GELU and SiLU) activation functions, and common loss functions including mean squared error, binary cross-entropy, and Huber loss. The advantage of our approach is that it is much easier to impose conditions such as Lipschitz smoothness or continuity on the layer-wise functions as opposed to the overall loss function. Unlike prior work, we have *no requirements on the width, depth, initialization or dataset.* As we do not restrict the iterates to a bounded set, our analysis seamlessly accounts for the feature learning regime of neural networks, where the parameters may progress arbitrarily far from initialization, even towards a stationary point at infinity.

Our contributions can be summarized as follows.

- We explicitly derive the *double polynomial smoothness* of feedforward neural networks and their loss functions, proving that for an L-layer neural network, the change in gradient is upper bounded by the change in parameter space, multiplied by a polynomial of the parameter norms of both endpoints of degree $2L - 2$.
- We establish the first convergence rate of gradient descent on neural networks with no conditions on the initialization, dataset, depth or width. We prove that gradient descent converges at rate $O(1/T^{1/L})$, provided that the learning rate is small enough with respect to the parameter norm and loss value at each step.

2 Related Work

Convergence to global minima of neural network loss functions. Due to the enormous success of neural networks, a large body of work focuses on identifying circumstances under which gradient

[1] In particular, we use "double" to highlight the dependence on the parameter norms at *both* endpoints, and also to distinguish from the use of the term "polynomial smoothness" in adjacent research fields such as the analysis of Approximate Message Passing algorithms (Nguyen & Kim, 2023).

descent converges to *global minima* of neural network loss functions, where the resulting model achieves zero error on the training set. These analyses typically rely on a patchwork of assumptions on the model dimension, initialization or dataset in order to demonstrate that some gradient dominance condition persists for all points in the gradient descent trajectory (although not globally), driving the loss to zero. Many such results hold in an over-parametrized regime where the hidden layers are extremely wide relative to the number of training samples (Li & Liang, 2018; Du et al., 2019; Allen-Zhu et al., 2019; Dana et al., 2025; Xu et al., 2023; Taheri et al., 2025). Still others consider *infinite-width* neural networks (Chizat et al., 2024; Barboni et al., 2026); in particular, one central framework is the neural tangent kernel (NTK) analysis, where as the width of the network goes to infinity, the gradient descent training process reduces to a kernel method. In these analyses, the iterates can be shown to remain in a bounded set close to initialization. They also often rely on random initialization; for example, the NTK analysis leverages the fact that an infinite-width neural network with random initializations converges to a Gaussian process by way of the Central Limit Theorem (Jacot et al., 2018). Finally, some works may require additional restrictions on the dataset; for example the works Arora et al. (2019a) and Chen et al. (2026) consider gradient descent for deep linear neural networks on whitened data, and Chatterjee (2026) analyzes a complementary setting where the neural network dimension is arbitrary, but the dimension of the input data is at least the number of data points. While these analyses provide important insights into how neural networks can achieve near-zero training loss under strict assumptions, they leave the broader behavior of gradient descent beyond such assumptions unexplained.

At the same time, a general analysis of the convergence rate of gradient descent to a *stationary point* for neural networks–independent of architecture, dataset or initialization—is conspicuously missing. A well-known issue regarding this setting is the lack of a global Lipschitz constant or bound on the Hessian norm as the iterates converge towards infinity (Sun, 2019). By contrast, if training remains within a compact set, the Hessian is necessarily bounded for all iterates, and the function is locally Lipschitz smooth. Accounting for the potentially unbounded behavior of the parameters is not just a technical challenge; it is critical to understanding the *feature learning* regime of neural networks, where the parameters change significantly from initialization as the model learns feature representations of the data. In contrast, it has been shown that the lazy training regime, where iterates remain close to initialization, leads to underperforming models and does not fully explain the power of deep learning (Chizat et al., 2019; Woodworth et al., 2020; Seleznova & Kutyniok, 2022).

Comparison with Wang (2026). The work Wang (2026) first observed that neural networks satisfy a polynomial smoothness condition that can be conveniently derived from recursively relating the Lipschitz properties of the activation functions at each layer. They prove that the loss function satisfies *polynomial generalized smoothness*, a similar but less specific version of double polynomial smoothness. However, they only utilize this condition to uniformly bound the Lipschitz constant on a ball of radius R around a stationary point, with the constant depending on R, the parameter norm of the stationary point, and the maximum gradient norm over the ball. Setting the learning rate small enough with respect to this constant, they prove *asymptotic* convergence to a stationary point as long as the neural network is sufficiently overparametrized, the data distribution is nondegenerate, and the (mean squared error) loss function satisfies a local dissipative condition. Critically, they require that the width of the neural network exceeds the number of training data samples multiplied by the data dimension, which is looser than some NTK assumptions but still highly unrealistic.

In contrast, our analytical framework has *no* conditions on the network width or dataset, and it applies to a broad range of loss functions. Instead of constraining the parameters to a bounded set, we carefully track the potentially unbounded growth of the parameter norms—corresponding to the *dynamic* local Lipschitz constant—to achieve convergence to a stationary point even if at infinity. Key aspects of our analysis rely on the explicit smoothness expression we derive, including the specific *degree* of the polynomial in terms of L, which enables us to establish the non-asymptotic $O(1/T^{1/L})$ convergence rate. In Appendix B, we further compare and contrast the works and discuss in detail why the condition in Wang (2026) inhibits the same kind of convergence analysis.

Generalized Lipschitz smoothness. In the interest of developing more faithful analyses of the neural network setting, a recent line of work has targeted relaxations of the Lipschitz smoothness condition that can still yield convergence of gradient algorithms, also known as *generalized Lipschitz smoothness* or *non-uniform smoothness*. The work Zhang et al. (2020) proposes the (L_0, L_1)-

smoothness condition, where the Hessian norm is bounded by an affine function of the gradient norm, to explain the performance of gradient clipping. This condition was then extended to polynomial functions (Wilson et al., 2019; Yang et al., 2025) and then general nondecreasing functions of the gradient norm (Li et al., 2023). Additional works have considered bounding the Hessian norm by polynomial functions of the objective value (Vaswani et al., 2026; Mei et al., 2021). However, these generalized conditions are typically *empirically* validated and are only mathematically satisfied by some special cases of neural networks—for example, two-layer neural networks where one layer is frozen (Vaswani et al., 2026), two-layer neural networks with a strictly increasing activation function on linearly separable data (Taheri & Thrampoulidis, 2023), or deep linear neural networks where the weights remain in a "strongly balanced" subspace at all times (Alimisis et al., 2026).

A related area of research is *relative smoothness*, motivated by mirror descent analysis, where the Hessian norm is upper bounded by the Hessian of some convex reference function that is used to map iterates to and from a dual space (Bauschke et al., 2017; Lu et al., 2018). In particular, the works (Lu et al., 2018; Fatkhullin & He, 2024) identify that a two-layer linear neural network satisfies a relative smoothness condition where the loss Hessian is bounded by a quadratic term in the parameter norms. This condition is close to the double polynomial smoothness presented in our work. However, they only address mirror descent using this specific reference function, not vanilla gradient descent, and they do not consider general neural networks with nonlinear activation functions.

Finally, some works attempt to circumvent issues with global smoothness by considering separately the local conditions around the gradient descent trajectory. The work Mishkin et al. (2024) proposes *directional smoothness*, where the gradient variation depends solely on the local conditions on the optimization path. However, this requires solving implicit equations to obtain the learning rate, and the existing analysis only applies to convex functions. The work Fox et al. (2026) introduces the notion of "glocal" smoothness, where the loss function may have a prohibitively large *global* Lipschitz constant but a smaller *local* Lipschitz constant. This enables the analysis of algorithms on more complex landscapes, but still excludes neural networks, which lack a global Lipschitz constant over an unbounded domain. Finally, the work Berahas et al. (2024) considers the broad scenario where one has access to a local first-order smoothness oracle that bounds the local Lipschitz smoothness constant within a ball of some radius R. However, a key challenge of this framing lies in choosing R at each step such that it includes the next iterate. In contrast, the double polynomial smoothness derived in this work holds *globally* for any two points regardless of their distance apart.

Matrix factorization and low-rank fine-tuning. Prior works have observed the close connection between matrix factorization and neural network training (Sun, 2019; Xiong et al., 2024). In particular, we build on the convergence proof structure of the Low-Rank Adaptation (LoRA) algorithm (Mu & Klabjan, 2026), which can be reinterpreted as training a specific two-layer linear neural network. However, the general neural network setting is significantly more complex due to the nonlinearity of the activation functions and the recursive relationships generated by the number of layers.

3 SETUP

Let $\odot$ denote the Hadamard product and $\|\cdot\|_F$ denote the Frobenius norm, where for the real matrix $X \in \mathbb{R}^{m\times n}$, we have $\|X\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2}$. We also write $\|\cdot\|$ for $\|\cdot\|_F$ throughout. We denote $\langle A, B\rangle_F = \mathrm{tr}(A^T B)$.

We consider empirical risk minimization on a dataset of n training samples $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Let the constant $c_x = \max\{1, \max_{i\in\{1,\dots,n\}} \|x_i\|_\infty\}$ represent an upper bound on any individual component of any input data, such that for all i, we have $\|x_i\| \le c_x d^{1/2}$.

We model the data with f, a feedforward neural network. Denote by $\sigma_\ell : \mathbb{R} \to \mathbb{R}$ the activation function at the ℓth layer, and we overload notation such that for $X \in \mathbb{R}^{n\times m}$, $\sigma_\ell(X) \in \mathbb{R}^{n\times m}$ denotes the element-wise application of the activation function. Then an L-layer feedforward neural network f, predicting on data sample x, is parametrized by L weight matrices $w_1, w_2, ..., w_L$ as

$$f(x; w_1, w_2, ..., w_L) = \sigma_L(w_L \sigma_{L-1}(w_{L-1}\sigma_{L-2}(...w_2\sigma_1(w_1 x)))),$$

where $L \geq 2$, $w_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, $d_0 = d$, and $d_L = 1$. We also denote the width of the neural network as $d_{max} = \max\{d_0, ..., d_L\}$. For the following analysis, we assume that the activation functions are linearly bounded, Lipschitz continuous, and Lipschitz smooth.

Assumption 3.1. For all ℓ, the activation function $\sigma_\ell : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and linearly bounded such that for all $w \in \mathbb{R}$, we have

$$|\sigma_\ell(w)| \leq c_0 + c_1|w|. \tag{1}$$

Moreover, σ_ℓ is Lipschitz continuous and Lipschitz smooth with nonnegative constants c_2, c_3 such that for all $w_1, w_2 \in \mathbb{R}$, we have

$$|\sigma_\ell(w_1) - \sigma_\ell(w_2)| \leq c_2|w_1 - w_2|, \qquad |\sigma'_\ell(w_1) - \sigma'_\ell(w_2)| \leq c_3|w_1 - w_2|. \tag{2}$$

Assumption 3.1 also implies that $|\sigma'_\ell(w)| \leq c_2$ for all $w \in \mathbb{R}$. To simplify the algebra, we state our result in terms of $c_{max} = \max\{c_0, c_1, c_2, c_3, 1\}$, where $c_{max} = 1$ for standard tanh, sigmoid, linear, and softplus functions, and c_{max} is slightly greater than 1 for GELU and SiLU functions (see Appendix C.11).

We construct a block matrix W with the weight matrices $w_1, ..., w_L$ on the diagonal as follows,

$$W = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_L \end{bmatrix}, \quad W \in \mathbb{R}^{D_2 \times D_1}, \quad \text{where } D_1 = \sum_{\ell=0}^{L-1} d_\ell, D_2 = \sum_{\ell=1}^{L} d_\ell.$$

For each layer ℓ, define the matrices $e_{\ell 1} \in \mathbb{R}^{d_\ell \times D_2}$ and $e_{\ell 2} \in \mathbb{R}^{D_1 \times d_{\ell-1}}$ as follows

$$e_{\ell 1} = [0_{d_\ell \times d_1} \quad 0_{d_\ell \times d_2} \quad \dots \quad 0_{d_\ell \times d_{\ell-1}} \quad I_{d_\ell \times d_\ell} \quad 0_{d_\ell \times d_{\ell+1}} \quad \dots \quad 0_{d_\ell \times d_L}],$$
$$e_{\ell 2} = [0_{d_{\ell-1} \times d_0} \quad 0_{d_{\ell-1} \times d_1} \quad \dots \quad 0_{d_{\ell-1} \times d_{\ell-2}} \quad I_{d_{\ell-1} \times d_{\ell-1}} \quad 0_{d_{\ell-1} \times d_\ell} \quad \dots \quad 0_{d_{\ell-1} \times d_{L-1}}]^T.$$

We define the linear operator $E_\ell[W]$ that extracts the ℓth diagonal block from W, recovering w_ℓ, as well as its corresponding adjoint operator $E^*_\ell[M]$ for $M \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$

$$E_\ell[W] = e_{\ell 1} W e_{\ell 2} = w_\ell, \qquad E^*_\ell[M] = e^T_{\ell 1} M e^T_{\ell 2}. \tag{3}$$

Now we can write f in a reparametrized form F as a function of $W \in \mathbb{R}^{D_2 \times D_1}$. We have

$$f(x_i; w_1, .., w_L) = F(x_i; W) = \sigma_L(E_L[W]\sigma_{L-1}(...E_2[W]\sigma_1(E_1[W]x_i))). \tag{4}$$

Let $\hat{y}_i = F(x_i; W)$ represent the prediction on the sample (x_i, y_i). We define the individual loss on the output data y_i as $J_i : \mathbb{R} \to \mathbb{R}$. Then the overall loss function $\mathcal{L} : \mathbb{R}^{D_2 \times D_1} \to \mathbb{R}$ is given as

$$\mathcal{L}(W) = \frac{1}{n} \sum_{i=1}^{n} J_i(\hat{y}_i) = \frac{1}{n} \sum_{i=1}^{n} J_i(F(x_i; W)). \tag{5}$$

We require a mild smoothness assumption on J_i, which is satisfied by common loss functions such as mean squared error (MSE), binary cross-entropy (BCE), and Huber loss (see Appendix C.11).

Assumption 3.2. Let $c_J \geq 1$. The individual loss J_i is nonnegative and c_J-Lipschitz smooth in the model outputs, such that for $u, v \in \mathbb{R}$ representing different model outputs, we have $|J'_i(u) - J'_i(v)| \leq c_J|u - v|$.

For MSE, $J_i(u) = (u - y_i)^2$, and $J'_i(u) = 2(u - y_i)$, so it satisfies this assumption with $c_J = 2$. The BCE loss function also satisfies Lipschitz smoothness with respect to the *logit* outputs before the softmax function. Specifically, for a model where the last activation function is linear, we define the BCE loss on the ith point as (see also, Appendix C.11)

$$J_i(u) = -\left[y_i \log(\frac{1}{1 + e^{-u}}) + (1 - y_i) \log(1 - \frac{1}{1 + e^{-u}})\right] = -uy_i + \log(1 + e^u).$$

When we restrict $y_i \in [0, 1]$, J_i is nonnegative. The absolute value of the second derivative $J''_i(u)$ is upper bounded by $1/4$, so BCE satisfies the assumption and we can choose $c_J = 1$. In addition, multi-class cross entropy also satisfies Lipschitz smoothness with respect to multiple logits and may

also be covered by our theory; however, the model gradient would be a tensor rather than a matrix, which would significantly complicate the notation. We therefore leave this case for future work.

We minimize $\mathcal{L}(W)$ using the gradient descent algorithm,

$$W_{t+1} = W_t - \eta_t \nabla \mathcal{L}(W_t), \tag{6}$$

where $\nabla \mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} J_i'(F(x_i; W)) \cdot \nabla_W F(x_i; W)$. The evolution of the block diagonal elements of W under gradient descent matches that of the original weight matrices $w_1, ..., w_L$, while the off diagonal elements exhibit zero change because they do not affect the value of F or $\mathcal{L}$.

4 Analyses

Broadly speaking, we aim to show that for any W_1, W_2, we can bound $\|\nabla \mathcal{L}(W_2) - \nabla \mathcal{L}(W_1)\|$ relative to $\|W_2 - W_1\|$. We first relate the change in loss gradient to the change in model function value and change in model gradient, as stated in Lemma 4.1.

Lemma 4.1. *For $W_1, W_2 \in \mathbb{R}^{D_2 \times D_1}$, we have*

$$\begin{aligned}\|\nabla \mathcal{L}(W_2) - \nabla \mathcal{L}(W_1)\| \leq & \frac{1}{n}\sum_{i=1}^{n} c_J |F(x_i; W_2) - F(x_i; W_1)| \|\nabla_W F(x_i; W_2)\| \\ & + \frac{1}{n}\sum_{i=1}^{n} |J_i'(F(x_i; W_1))| \|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\|.\end{aligned}$$

Now we explicitly compute $\nabla_W F(x_i; W)$ for some x_i. Defining $h_0(W) = x_i$, we recursively define the intermediate variables $z_\ell(W) \in \mathbb{R}^{d_\ell}$ and $h_\ell(W) \in \mathbb{R}^{d_\ell}$ for $\ell = 1, ..., L$ as follows

$$z_\ell(W) = E_\ell[W] h_{\ell-1}(W), \qquad h_\ell(W) = \sigma_\ell(z_\ell(W)). \tag{7}$$

Then we have $F(W) = h_L(W) = \sigma_L(z_L(W))$. We also define the variable $\delta_\ell \in \mathbb{R}^{d_\ell}$ as

$$\delta_L(W) = \sigma_L'(z_L(W)), \qquad \delta_\ell(W) = \sigma_\ell'(z_\ell(W)) \odot (E_{\ell+1}[W]^T \delta_{\ell+1}(W)). \tag{8}$$

Lemma 4.2. *For some input data x_i, the gradient of F with respect to W is given as*

$$\nabla_W F(x_i; W) = \sum_{\ell=1}^{L} E_\ell^*[\delta_\ell(W) h_{\ell-1}(W)^T].$$

We decompose the change in $\nabla_W F(x_i; W)$ in terms of changes in δ_ℓ and h_ℓ, which can be defined recursively in terms of other layers. Because of the Lipschitz properties of each σ_ℓ from Assumption 3.1, these terms can also be bounded by $\|W_2 - W_1\|$. Collecting these terms together allows us to establish the *double polynomial smoothness* of F, where the change in gradient is upper bounded by $\|W_2 - W_1\|$, multiplied by a polynomial of the endpoints $\|W_1\|$ and $\|W_2\|$.

Lemma 4.3. *For all W_1, $W_2 \in \mathbb{R}^{D_2 \times D_1}$, and x_i, we have*

$$\|F(x_i; W_2) - F(x_i; W_1)\| \leq c_x d_{max}^{1/2} c_{max}^L \|W_2 - W_1\| \sum_{j=0}^{L-1} \sum_{k=0}^{L-j-1} \|W_1\|^j \|W_2\|^k.$$

Lemma 4.4. *For all W_1, $W_2 \in \mathbb{R}^{D_2 \times D_1}$, and x_i, we have*

$$\|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| \leq 4 c_x^2 d_{max} c_{max}^{3L} L^3 \|W_2 - W_1\| \sum_{j=0}^{2L-2} \|W_1\|^j \sum_{k=0}^{2L-j-2} \|W_2\|^k.$$

By plugging Lemmas 4.3 and 4.4 into Lemma 4.1, we have the following bound establishing the double polynomial smoothness of $\mathcal{L}$.

Lemma 4.5. *For all $W_1, W_2 \in \mathbb{R}^{D_2 \times D_1}$, we have*

$$\|\nabla \mathcal{L}(W_2) - \nabla \mathcal{L}(W_1)\| \leq 8 c_J c_x^2 d_{max} c_{max}^{3L} L^3 \|W_2 - W_1\| (1 + \mathcal{L}(W_1)^{1/2}) \sum_{j=0}^{2L-2} \|W_1\|^j \sum_{k=0}^{2L-j-2} \|W_2\|^k.$$

Lemma 4.5 also includes a factor of $1 + \mathcal{L}(W_1)^{1/2}$ in addition to the polynomials of $\|W_1\|$, $\|W_2\|$. We note that $\mathcal{L}(W_1)$ can also be upper bounded in terms of a polynomial of $\|W_1\|$, but leaving it in this form allows us to leverage descent in $\mathcal{L}$ over time. In fact, due to this factor of $\mathcal{L}(W_1)$, the right hand side of Lemma 4.5 is *asymmetric* in its dependence on W_1 and W_2. Even if we assume twice-differentiability, this kind of asymmetric smoothness condition is not equivalent to any pointwise bound on the Hessian. This departs from many prior works, such as analyses of (L_0, L_1)-smoothness, which assume a direct equivalence between smoothness and Hessian bounds.

Lemma 4.5 directly leads to the following descent lemma. In particular, after integrating over the segment between W_1 and W_2, the polynomial terms in $\|W_2\|$ translate to higher order terms of $\|W_2 - W_1\|$. This differs from the classic Lipschitz smoothness descent lemma, which only contains a quadratic term of $\|W_2 - W_1\|$.

Lemma 4.6. *For all W_1 and $W_2 \in \mathbb{R}^{D_2 \times D_1}$, we have*

$$\begin{aligned}\mathcal{L}(W_2) - \mathcal{L}(W_1) \leq &\langle \nabla \mathcal{L}(W_1), W_2 - W_1 \rangle_F \\ &+ C\Big[(1 + \mathcal{L}(W_1)^{1/2}) \sum_{j=0}^{2L-2} \|W_1\|^j\Big] \sum_{k=0}^{2L-2} \frac{1}{k+2} \|W_2 - W_1\|^{k+2},\end{aligned}$$

where $C = 2^{2L+2} c_J c_x^2 d_{max} c_{max}^{3L} L^4$.

Applying this lemma to the gradient descent iterates W_t and W_{t+1} (6) yields descent in the loss function value as long as η_t is chosen appropriately. The formula for η_t only requires access to the current parameter norm $\|W_t\|$ and loss value $\mathcal{L}(W_t)$ and is therefore easily computable without any subroutines or implicit solutions.

Lemma 4.7. *For one step of gradient descent $W_{t+1} = W_t - \eta_t \nabla \mathcal{L}(W_t)$, if η_t satisfies*

$$\eta_t = \frac{1}{\rho(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i}, \tag{9}$$

where $\rho = 2^{2L+3} c_J c_x^2 d_{max} c_{max}^{3L} L^5$, then the function value descends in one step,

$$\mathcal{L}(W_{t+1}) - \mathcal{L}(W_t) \leq -\frac{\eta_t}{2L} \|\nabla \mathcal{L}(W_t)\|^2,$$

and for T steps we obtain the bound

$$\sum_{t=0}^{T-1} \eta_t \|\nabla \mathcal{L}(W_t)\|^2 \leq 2L \mathcal{L}(W_0). \tag{10}$$

To show convergence of $\|\nabla \mathcal{L}(W_t)\|^2$ to zero, we require that the learning rate (9) not decay too quickly. By Lemma 4.7, we can bound $\mathcal{L}(W_t) \leq \mathcal{L}(W_0)$. It remains to control the growth of $\sum_{i=0}^{2L-2} \|W_t\|^i$, which is dominated by the leading term $\|W_t\|^{2L-2}$. We have for $N = 2L - 2$,

$$\|W_T\|^N \leq \|W_0\|^N + N \sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{t=0}^{T-1} \sum_{k=2}^{N} \binom{N}{k} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k.$$

By (10), it can be shown that the last term is bounded in terms of the constant $2L\mathcal{L}(W_0)$. However, the growth of $\sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\|$ depends on the value of N. For two-layer neural networks, $L = 2$ and $N = 2$, and we can directly use (10) to show that this term grows as $O(\sqrt{T})$, yielding an $O(1/\sqrt{T})$ convergence rate. However, for deeper neural networks $L > 2$, we have to bound the sum via Cauchy-Schwartz as follows,

$$\sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| = \sum_{t=0}^{T-1} \eta_t^{1/N - 1/2} \eta_t^{1/2} \|\nabla \mathcal{L}(W_t)\| \leq \Big(\sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{N}} \Big)^{1/2} \Big(\sum_{t=0}^{T-1} \eta_t \|\nabla \mathcal{L}(W_t)\|^2 \Big)^{1/2}.$$

Since $(\sum_{t=0}^{T-1} \eta_t \|\nabla \mathcal{L}(W_t)\|^2)^{1/2}$ can also be bounded via (10), it remains to determine the growth of $(\sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{N}})^{1/2}$. We show by induction that $(\sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{N}})^{1/2}$ and therefore $\|W_T\|^{2L-2}$ grows as $O(T^{\frac{L-1}{L}})$, leading to an overall $O(1/T^{1/L})$ convergence rate for the minimum squared gradient norm. We state our main result as follows.

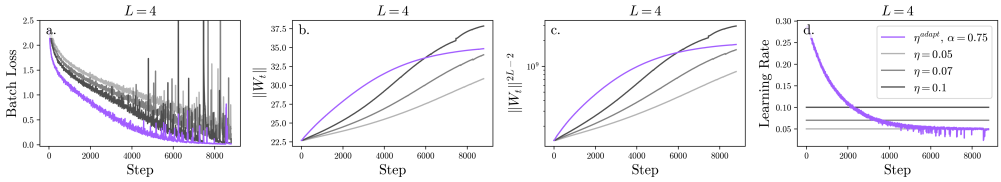

Figure 1: Adaptive vs. constant learning rate for standard neural network with $L = 4$.

Theorem 4.8. *Consider the loss function $\mathcal{L}$ (5) of an L-layer neural network (4). Suppose Assumptions 3.1 and 3.2 hold. Then after T steps of gradient descent (6), with η_t set as (9), we have*

$$\min_{t=0,\dots,T-1} \|\nabla \mathcal{L}(W_t)\|^2 = O\Big(\frac{1}{T^{1/L}}\Big).$$

Corollary 4.9. *If we have $\|W_t\| \leq R$ for all t, then after T steps of gradient descent with the constant learning rate $\eta = \frac{1}{\rho(1+\mathcal{L}(W_0)^{1/2})\sum_{i=0}^{2L-2} R^i}$ we have $\min_{t=0,\dots,T-1}\|\nabla \mathcal{L}(W_t)\|^2 = O(1/T)$.*

In the above statements, the $O(\cdot)$ notation hides polynomial dependence on c_J, c_x, d_{max}, c_{max}, $\mathcal{L}(W_0)$, and $\|W_0\|$, and exponential dependence on L. Theorem 4.8 establishes the first upper bound of the convergence rate of gradient descent on neural network loss functions of any depth or width, for any initialization or dataset. Moreover, our analysis does not rely on requiring that the parameters remain within a bounded set; we demonstrate that gradient descent converges even if $\|W_T\|$ grows to infinity, capturing the feature learning regime of neural networks. As the number of layers L increases, the local geometry (controlled by $\|W_t\|^{2L-2}$) grows more sharp, forcing the learning rate to decay faster and slowing convergence. Whether this phenomenon represents a fundamental limitation or is an artifact of the analysis remains an open question.

5 Experiments

We investigate practical implications of our theory by training feedforward neural networks of varying depths on the CIFAR-10 image classification dataset (Krizhevsky, 2009). We fix the width of the internal layers as 512, and we use the GELU activation function due to its close similarity to ReLU, a popular activation function. We use minibatch gradient descent with batch size 512 on the cross-entropy loss function. For additional experimental details and results, including the anonymous Github repository, see Appendix D.

First, we demonstrate that scaling the learning rate according to (9) leads to improved convergence. We define the following adaptive learning rate η^{adapt}, which scales with the parameter norm and batch loss value at each step, replacing the constants in (9) with a tunable hyperparameter α,

$$\eta_t^{adapt} = \frac{\alpha \|W_0\|^{2L-2}}{(1+\mathcal{L}(W_t)^{1/2})\sum_{j=0}^{2L-2}\|W_t\|^j}.$$

Figure 1 displays the loss and parameter norms over the training steps for constant and adaptive learning rates, for a neural network of depth $L = 4$ (for other values of L, see Figure 5 in Appendix D). For all cases, η_t^{adapt} displays faster and more stable convergence than constant learning rates of comparable values. In particular, η_t^{adapt} adapts to the local geometry of the loss function by starting at a large value and decaying over time, which accelerates the iterates away from initialization.

Second, we examine whether increasing the number of layers in a neural network indeed slows down convergence, as suggested by Theorem 4.8. For standard neural networks, the effect of increasing depth is challenging to isolate because increasing the depth also increases the number of parameters, which may substantially change other factors such as the overfitting of the model or the size of the parameter norms. However, our convergence analysis actually applies to a more general class of problems than the standard neural network structure, as the specific choice of the projection operators does not matter to the proof as long as they satisfy certain contraction properties (Lemma C.6). This allows us to investigate the impact of depth for a *shared-weights* neural network, where

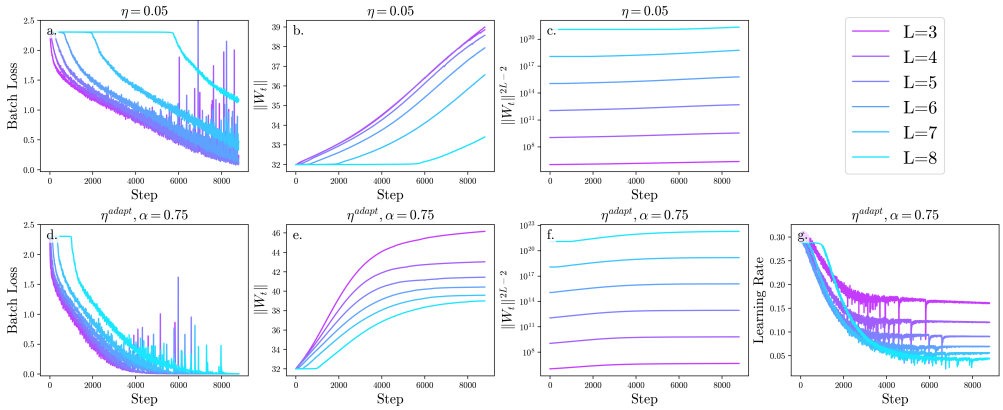

Figure 2: Training of shared-weights neural networks of varying L with constant (top row) and adaptive (bottom row) learning rates for 100 epochs.

each layer shares roughly the same weight matrix. We construct such a neural network where for d-dimensional input data, the input layer is $512 \times d$, and all other layers are a fixed 512×512 subset of the input layer weights, except for the output layer which is a 1×512 subset. This model has a fixed size and number of parameters even as the number of layers increases.

Figure 2 displays the results of training the shared-weights models for increasing L. As in Figure 1, the adaptive learning rate displays faster and more stable convergence than constant learning rates. Moreover, for both types of learning rates, we observe the convergence is indeed consistently slower for larger L. For example, the $L = 8$ case converges the slowest though its η^{adapt} is higher for much of the training period. Our results suggest that the rate dependence on L in Theorem 4.8 is indeed fundamental, at least for the general class of models we consider. If a tighter convergence rate for standard neural networks were shown to be *independent* of L, it would be solely due to the structure of the specific projection operators.

The exception to the above trend is $L = 2$, for which the loss decayed quickly early on but plateaued above all other loss curves for the rest of training. We hypothesize that this is because the iterates are converging to a different stationary point. Investigating this further, we extend the training period from 100 epochs to 1,000 epochs for $L = 2$, with the results shown in Figure 3. While the iterates in Figure 2 appear to converge to a global minima within the training period, the iterates in Figure 3 appear to be converging to a stationary point at infinity, supported by the unbounded growth of $\|W_t\|$ in 3b even as the loss appears to have converged. This reinforces the importance of analyzing gradient descent on neural networks without assuming the iterates remain in a bounded set.

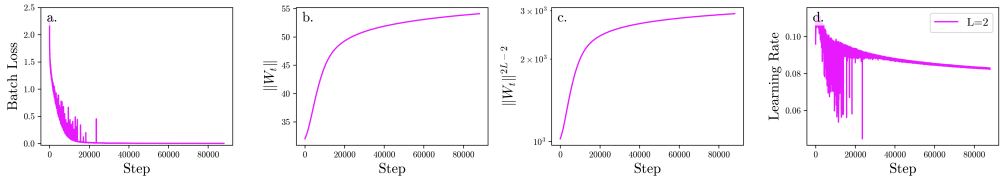

Figure 3: Extended training (1000 epochs) for $L = 2$ shared-weights neural network.

6 Conclusion

We establish the first convergence rate of gradient descent on L-layer neural networks with no requirements on the depth, width, initialization, or dataset, proving that with an appropriate learning rate, the minimum squared gradient norm converges at rate $O(1/T^{1/L})$. Our analysis relies on precisely characterizing the double polynomial smoothness of neural network loss functions, which holds for a broad range of popular activation functions and loss functions. By accounting for the potentially unbounded growth of the parameters away from initialization, our work bypasses the restrictive overparameterization and lazy-training limits that fail to capture practical feature learning.

7 AI USE STATEMENT

In this work, we used generative AI tools to assist with developing conceptual frameworks, formulating mathematical claims, writing and checking of proofs and the refining of mathematical statements. We have not used generative AI tools for providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing research experiments, cleaning or processing data, or interpreting results, and generating synthetic datasets, translation, and data analysis are not applicable to this work. Additionally, we used generative AI tools for brainstorming, searching existing literature, implementing methods and creating/editing software code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sahin Ahmed. How to choose the right learning rate in deep learning (with PyTorch), February 2025. URL https://medium.com/@sahin.samia/how-to-choose-the-right-learning-rate-in-deep-learning-with-pytorch-690de782b405.

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective, 2026.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019b.

Raphaël Barboni, Maarten V. de Hoop, Takashi Furuya, and Gabriel Peyré. Training infinitely deep and wide transformers, 2026.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, May 2017. ISSN 0364-765X. doi: 10.1287/moor.2016.0817.

Albert S. Berahas, Lindon Roberts, and Fred Roosta. Non-uniform smoothness for gradient descent. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Sourav Chatterjee. Convergence of gradient descent for deep neural networks, 2026.

Kun Chen, Dachao Lin, and Zhihua Zhang. A non-local convergence analysis of gradient flow for deep linear networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026. doi: 10.1109/TPAMI.2026.3704447.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. *On lazy training in differentiable programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks, 2024.

Léo Dana, Loucas Pillaud-Vivien, and Francis Bach. Convergence of the gradient flow for shallow reLU networks on weakly interacting data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.12.012. Special issue on deep reinforcement learning.

Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general Bregman divergence. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3493–3501. PMLR, 02–04 May 2024.

Curtis Fox, Aaron Mishkin, Sharan Vaswani, and Mark Schmidt. Glocal smoothness: Line search and adaptive sizes can help in theory too! *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. Expert Certification.

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods, 2024.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL `http://arxiv.org/abs/1606.08415`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8580–8589, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. URL `https://www.cs.toronto.edu/~kriz/cifar.html`.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 8168–8177, Red Hook, NY, USA, 2018. Curran Associates Inc.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: 10.1137/16M1099546.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7555–7564. PMLR, 18–24 Jul 2021.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 14810–14848. Curran Associates, Inc., 2024. doi: 10.52202/079017-0473.

Siqiao Mu and Diego Klabjan. On the convergence rate of lora gradient descent, 2026.

Dang Qua Nguyen and Taejoon Kim. On the stability of approximate message passing with independent measurement ensembles, 2023.

Mariia Seleznova and Gitta Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 868–895. PMLR, 16–19 Aug 2022.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017. doi: 10.1109/WACV.2017.58.

Ruoyu Sun. Optimization for deep learning: theory and algorithms, 2019.

Hossein Taheri and Christos Thrampoulidis. Fast convergence in learning two-layer neural networks with separable data, 2023.

Hossein Taheri, Christos Thrampoulidis, and Arya Mazumdar. Sharper guarantees for learning neural network classifiers with gradient methods. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 11836–11847, 2025.

Sharan Vaswani, Yifan Sun, and Reza Babanezhad. Convergence of steepest descent and Adam under non-uniform smoothness, 2026.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Empirical limitations of the NTK for understanding scaling laws in deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Yuqing Wang. Convergence of gradient descent for general neural network architectures beyond the NTK regime, 2026.

Ashia Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent: Fast optimization of smooth functions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020.

Nuoya Xiong, Lijun Ding, and Simon Shaolei Du. How over-parameterization slows down gradient descent in matrix sensing: The curses of symmetry and initialization. In *The Twelfth International Conference on Learning Representations*, 2024.

Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2262–2284. PMLR, 25–27 Apr 2023.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2022.

Yufeng Yang, Erin E. Tripp, Yifan Sun, Shaofeng Zou, and Yi Zhou. Adaptive gradient normalization and independent sampling for (stochastic) generalized-smooth optimization. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

A Additional Discussion

Our results open up a number of questions regarding the behavior of gradient descent on neural networks. The most natural question is regarding the tightness of the bounds, and obtaining a lower bound on the convergence rate. In particular, if the $O(1/T^{1/L})$ bound is tight, then the convergence rate of gradient descent is slowed as the number of layers increases. It would also open up a performance gap that might explain why other optimization algorithms, like ADAM, outperform gradient descent on neural networks. In addition, gradient descent on Lipschitz smooth functions is famously *dimension-free*, but our analysis includes polynomial dependence on the network width d_{max} and exponential dependence on the number of layers L. Determining whether these dependencies can be removed is an important future direction. Finally, although we significantly relax the assumptions of prior work, this classical framework still heavily relies on some kind of smoothness—in plain words, that shrinking in parameter space continuously maps to shrinking in the gradient. This leaves the behavior of gradient descent on ReLU neural networks largely unexplained.

B Further Comparison with Wang (2026)

The work Wang (2026) proposes polynomial generalized smoothness, where for a function ϕ and weights W_1, W_2, there exists a polynomial $P(\cdot)$ such that

$$\|\nabla\phi(W_2) - \nabla\phi(W_1)\| \leq P(\max\{\|W_1\|, \|W_2\|\})\|W_2 - W_1\|$$

They demonstrate this property is generally preserved under addition, multiplication, and composition. For neural networks, by repeatedly composing the activation functions, they show that there exists $P(\cdot)$ such that

$$\|\nabla\mathcal{L}(W_2) - \nabla\mathcal{L}(W_1)\| \leq P(\max\{\|W_1\|, \|W_2\|\})\|W_2 - W_1\|, \tag{11}$$

which leads to the following descent lemma,

$$\mathcal{L}(W_2) \leq \mathcal{L}(W_1) + \langle\nabla\mathcal{L}(W_1), W_2 - W_1\rangle_F + \frac{1}{2}P(\max\{\|W_1\|, \|W_2\|\})\|W_2 - W_1\|^2. \tag{12}$$

They do not derive the specific form of P beyond demonstrating that the exponents are natural numbers with some dependence on L, and that the number of terms is also some unspecified polynomial of L (see Appendix B.2 of Wang (2026)).

While the above properties share similarities with the conditions derived in our work, they have the following two limitations. First, in order to dynamically adjust the learning rate η_t to the local smoothness via (12), one needs to set η_t small enough with respect to $P(\max\{\|W_t\|, \|W_{t+1}\|\})$ which depends implicitly on information at the $t+1$ step that is not readily available. Instead, Wang (2026) uses tools from measure theory and dynamical systems to upper bound the Lipschitz constant c within a ball of radius R as

$$c = P(\|W^*\| + R + \max_{W \in B_{W^*(R)}} \|\nabla\mathcal{L}(W)\|),$$

where W^* is a stationary point of the function and R depends on the local dissipative condition of the function. This is a static bound over the whole ball that also depends on the maximum gradient norm on the whole ball. They pick a *constant* learning rate, inversely proportional to c, to achieve *asymptotic convergence* to a stationary point. The work also relies on a a very wide neural network where $d_{L-1} > nd$, which never holds in practice, to leverage a gradient dominance property.

In contrast, although we show that the local smoothness *does* depend on both $\|W_1\|$ and $\|W_2\|$, when deriving the descent lemma (Lemma 4.6) we can cleanly integrate over the line segment between W_1 and W_2, leaving behind only terms dependent on W_1. We can therefore dynamically set η_t in terms of $\|W_t\|$ and $\mathcal{L}(W_t)$, which are easily accessible. Moreover, η_t does not depend on hard-to-compute values like the location of a stationary point or the maximum gradient norm over a ball.

The second limitation is that they do not derive the specific form of P satisfied by an L-layer neural network. Indeed, the generic Lipschitz assumptions considered in both Wang (2026) and our work lead to an explosion of terms that require careful tracking and bounding to achieve a simple and interpretable form. A key contribution of our analysis is decomposing the change in $\nabla_W F(x_i; W)$

and $\nabla\mathcal{L}(W)$ into difference terms of the components, bounding those in terms of $\|W_2 - W_1\|$, and combining the multiple sums cleanly to reduce the bound as much as possible. This leads to Lemma 4.5 which states (up to some constants)

$$\|\nabla\mathcal{L}(W_2) - \nabla\mathcal{L}(W_1)\| \leq (1 + \mathcal{L}(W_1)^{1/2})\tilde{P}(\|W_1\|, \|W_2\|)\|W_2 - W_1\|,$$

where

$$\tilde{P}(\|W_1\|, \|W_2\|) = \sum_{j=0}^{2L-2} \|W_1\|^j \sum_{i=0}^{2L-j-2} \|W_2\|^i.$$

A critical feature of our analysis is that reducing the power of the leading term of $\tilde{P}$ tightens the upper bound on the convergence rate of gradient descent, because if $\tilde{P}$ is of degree N, $\|W_t\|^N$ grows at rate $T^{N/(N+2)}$. By simplifying the bound to a polynomial of degree $N = 2L-2$, we show that $\|W_t\|^{2L-2}$ grows at rate $T^{\frac{L-1}{L}}$, leading to an overall $O(\frac{1}{T^{1/L}})$ convergence rate. This represents the first convergence rate of gradient descent on neural networks without any requirements on width, depth, initialization, or dataset.

C Proofs

C.1 Helper Lemmas

In this section, we include some basic or well-known theorems and identities required for the analysis.

Lemma C.1. *Let $v \in \mathbb{R}^d$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be a scalar function such that $|\sigma(x)| \leq c_0 + c_1|x|$ for all $x \in \mathbb{R}$, where $c_0, c_1 \geq 0$. If $\sigma(v)$ denotes the element-wise application of σ to v, then the Euclidean (L_2) norm of $\sigma(v)$ satisfies the inequality:*

$$\|\sigma(v)\|_2 \leq c_0\sqrt{d} + c_1\|v\|_2$$

Proof. We can show this using the triangle inequality.

$$\|\sigma(v)\| = \sqrt{\sum_{i=1}^{d} \sigma(v_i)^2} \leq \sqrt{\sum_{i=1}^{d} (c_0 + c_1|v_i|)^2} \leq \sqrt{\sum_{i=1}^{d} c_0^2} + \sqrt{\sum_{i=1}^{d} c_1^2|v_i|^2} \leq c_0\sqrt{d} + c_1\|v\|_2.$$

□

Theorem C.2. *(Binomial theorem) For any nonnegative integer n and scalars x, y we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Lemma C.3. *For vectors $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$ and matrix $M \in \mathbb{R}^{d_1 \times d_2}$, we have*

$$\langle a, Mb\rangle = \langle M^T a, b\rangle.$$

Proof.

$$\langle a, Mb\rangle = b^T M^T a = \langle b, M^T a\rangle.$$

□

Lemma C.4. *For vectors $a, b, c \in \mathbb{R}^d$, we have*

$$\langle a \odot b, c\rangle = \langle b, a \odot c\rangle.$$

Proof. Denote $diag(a)$ as the diagonal $d \times d$ matrix with a on the diagonal. Then

$$\begin{aligned} a \odot b =& diag(a)b, \\ \langle a \odot b, c\rangle =& \langle diag(a)b, c\rangle, \\ =& \langle b, diag(a)^T c\rangle, \\ =& \langle b, diag(a)c\rangle, \\ =& \langle b, a \odot c\rangle. \end{aligned}$$

□

Lemma C.5. *(Lp norms) For a sequence of scalars $a_1, .., a_n$, we have for $p \geq 2$,*

$$\big(\sum_{i=1}^{n} |a_i|^p\big)^{1/p} \leq \big(\sum_{i=1}^{n} |a_i|^2\big)^{1/2} \leq \sum_{i=1}^{n} |a_i| \leq \sqrt{n}\big(\sum_{i=1}^{n} |a_i|^2\big)^{1/2}.$$

Lemma C.6. *Let E_ℓ and E_ℓ^* be defined as in (3). For all $W \in \mathbb{R}^{D_2 \times D_1}$ and $M \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, we have*

$$\begin{aligned} \|E_\ell[W]\| \leq \|W\|, \\ \|E_\ell^*[M]\| \leq \|M\|. \end{aligned}$$

Proof. Since E_ℓ just extracts a $d_\ell \times d_{\ell-1}$ subset of W, the first statement holds. Since E_ℓ^* produces a larger $D_2 \times D_1$ matrix containing M and zero entries, the second statement holds. □

Lemma C.7. *Suppose a scalar function $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant c, such that for any $w_1, w_2 \in \mathbb{R}$, we have*

$$|\sigma(w_1) - \sigma(w_2)| \leq c|w_1 - w_2|.$$

Then the elementwise extension of σ to matrices is also Lipschitz continuous with the same constant, such that for $W, W' \in \mathbb{R}^{d_1 \times d_2}$, we have

$$\|\sigma(W) - \sigma(W')\|_F \leq c\|W - W'\|_F.$$

Proof.

$$\begin{aligned} \|\sigma(W) - \sigma(W')\|_F =& \sqrt{\sum_{i,j} (\sigma(W_{ij}) - \sigma(W'_{ij}))^2} \\ \leq& \sqrt{\sum_{i,j} c^2 (W_{ij} - W'_{ij})^2} \\ =& c\|W - W'\|_F. \end{aligned}$$

□

Lemma C.8. *Let $A, B \in \mathbb{R}^{m \times n}$ be real-valued matrices, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be an element-wise function such that $|\sigma(x)| \leq c_0 + c_1|x|$ for constants $c_0, c_1 \geq 0$ and for all $x \in \mathbb{R}$. Then we have*

$$\|\sigma(A) \odot B\|_F \leq (c_0 + c_1\|A\|_F)\|B\|_F$$

Proof. We have

$$\begin{aligned}
\|\sigma(A)\odot B\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|(\sigma(A)\odot B)_{ij}|^2} &= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|\sigma(A_{ij})B_{ij}|^2} \\
&\leq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}((c_0+c_1|A_{ij}|)|B_{ij}|)^2} \\
&\leq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(c_0|B_{ij}|+c_1|A_{ij}||B_{ij}|)^2} \\
&\leq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}c_0^2|B_{ij}|^2}+\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}c_1^2|A_{ij}|^2|B_{ij}|^2} \\
&\leq c_0\|B\|+c_1\|A\|\|B\|.
\end{aligned}$$

□

C.2 Proof of Lemma 4.1

Proof. We have $\nabla\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} J_i'(F(x_i;W))\cdot\nabla_W F(x_i;W)$. The change in gradient is as follows,

$$\begin{aligned}
\|\nabla\mathcal{L}(W_2)-\nabla\mathcal{L}(W_1)\| =& \|\frac{1}{n}\sum_{i=1}^{n}J_i'(F(x_i;W_2))\nabla_W F(x_i;W_2)-\frac{1}{n}\sum_{i=1}^{n}J_i'(F(x_i;W_1))\nabla_W F(x_i;W_1)\| \\
=& \|\frac{1}{n}\sum_{i=1}^{n}\Big[J_i'(F(x_i;W_2))\nabla_W F(x_i;W_2)-J_i'(F(x_i;W_1))\nabla_W F(x_i;W_2) \\
&+J_i'(F(x_i;W_1))\nabla_W F(x_i;W_2)-J_i'(F(x_i;W_1))\nabla_W F(x_i;W_1)\Big]\| \\
=& \|\frac{1}{n}\sum_{i=1}^{n}\Big[(J_i'(F(x_i;W_2))-J_i'(F(x_i;W_1)))\nabla_W F(x_i;W_2) \\
&+J_i'(F(x_i;W_1))(\nabla_W F(x_i;W_2)-\nabla_W F(x_i;W_1))\Big]\| \\
\leq& \|\frac{1}{n}\sum_{i=1}^{n}(J_i'(F(x_i;W_2))-J_i'(F(x_i;W_1)))\nabla_W F(x_i;W_2)\| \\
&+\|\frac{1}{n}\sum_{i=1}^{n}J_i'(F(x_i;W_1))(\nabla_W F(x_i;W_2)-\nabla_W F(x_i;W_1))\| \\
\overset{\text{Assumption 3.2}}{\leq}& \frac{1}{n}\sum_{i=1}^{n}c_J|F(x_i;W_2)-F(x_i;W_1)|\|\nabla_W F(x_i;W_2)\| \\
&+\frac{1}{n}\sum_{i=1}^{n}|J_i'(F(x_i;W_1))|\|\nabla_W F(x_i;W_2)-\nabla_W F(x_i;W_1)\|.
\end{aligned}$$

□

C.3 Proof of Lemma 4.2

Proof. For an element-wise function $\sigma(X)$, we have

$$d\sigma(X) = \sigma'(X)\odot dX.$$

So we have the following differentials.

$$\begin{aligned} dh_\ell =& \sigma'_\ell(z_\ell(W)) \odot dz_\ell \\ dz_\ell =& d(E_\ell[W])h_{\ell-1} + E_\ell[W]dh_{\ell-1} \\ =& E_\ell[dW]h_{\ell-1} + E_\ell[W]dh_{\ell-1} \\ dz_1 =& E_1[dW]x_i \end{aligned}$$

We have $\delta_L(W) = \sigma'_L(z_L(W))$, and $\delta_{\ell-1}(W) = \sigma'_{\ell-1}(z_{\ell-1}(W)) \odot (E_\ell[W]^T \delta_\ell(W))$. Then

$$dF = \langle \delta_L, dz_L \rangle,$$

and for $\ell = 2, .., L$,

$$\begin{aligned} \langle \delta_\ell(W), dz_\ell \rangle =& \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1} \rangle + \langle \delta_\ell(W), E_\ell[W]dh_{\ell-1} \rangle, \\ =& \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1} \rangle + \langle \delta_\ell(W), E_\ell[W](\sigma'_{\ell-1}(z_{\ell-1}(W)) \odot dz_{\ell-1}) \rangle, \\ \overset{\text{Lemma C.3}}{=}& \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1} \rangle + \langle E_\ell[W]^T \delta_\ell(W), \sigma'_{\ell-1}(z_{\ell-1}(W)) \odot dz_{\ell-1} \rangle \\ \overset{\text{Lemma C.4}}{=}& \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1} \rangle + \langle \sigma'_{\ell-1}(z_{\ell-1}(W)) \odot (E_\ell[W]^T \delta_\ell(W)), dz_{\ell-1}(W) \rangle \\ =& \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1} \rangle + \langle \delta_{\ell-1}(W), dz_{\ell-1} \rangle. \end{aligned}$$

For $\ell = 1$ we have

$$\langle \delta_1(W), dz_1 \rangle = \langle \delta_1(W), E_1[dW]h_0(W) \rangle,$$

which yields the following equation for dF:

$$\begin{aligned} dF =& \sum_{\ell=1}^{L} \langle \delta_\ell(W), E_\ell[dW]h_{\ell-1}(W) \rangle, \\ =& \langle \sum_{\ell=1}^{L} E^*_\ell[\delta_\ell(W)h_{\ell-1}(W)^T], dW \rangle. \end{aligned}$$

We therefore have the following formula for the gradient $\nabla_W F(x_i; W)$

$$\nabla_W F(x_i; W) = \sum_{\ell=1}^{L} E^*_\ell[\delta_\ell(W)h_{\ell-1}(W)^T].$$

□

C.4 Proof of Lemma 4.3

Proof.

$$\begin{aligned} \|F(x_i; W_2) - F(x_i; W_1)\| =& \|\sigma_L(z_L(W_2)) - \sigma_L(z_L(W_1))\|, \\ \leq& c_2 \|z_L(W_2) - z_L(W_1)\|, \\ \overset{\text{Lemma C.10}}{\leq}& c_x d_{max}^{1/2} c_{max}^L \|W_2 - W_1\| \sum_{j=0}^{L-1} \sum_{k=0}^{L-j-1} \|W_1\|^j \|W_2\|^k, \\ =& c_x d_{max}^{1/2} c_{max}^L \|W_2 - W_1\| \sum_{k=0}^{L-1} \sum_{j=0}^{L-k-1} \|W_1\|^j \|W_2\|^k. \end{aligned}$$

□

C.5 Proof of Lemma 4.4

Proof. We decompose the change in $\nabla_W F(x_i; W)$ in terms of the change in δ and h as follows

$$\begin{aligned}
\|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| &\leq \sum_{\ell=1}^{L} \|E^*_\ell[\delta_\ell(W_2)h_{\ell-1}(W_2)^T - \delta_\ell(W_1)h_{\ell-1}(W_1)^T]\|, \\
&\leq \sum_{\ell=1}^{L} \|\delta_\ell(W_2)h_{\ell-1}(W_2)^T - \delta_\ell(W_1)h_{\ell-1}(W_1)^T\|, \\
&\leq \sum_{\ell=1}^{L} [\|\delta_\ell(W_2)\|\|h_{\ell-1}(W_2) - h_{\ell-1}(W_1)\| + \|h_{\ell-1}(W_1)\|\|\delta_\ell(W_2) - \delta_\ell(W_1)\|].
\end{aligned}$$

We break up the sum as follows because $\|h_0(W_2) - h_0(W_1)\| = \|x_i - x_i\| = 0$, and to handle separately the bound on $\|\delta_L(W_2) - \delta_L(W_1)\|$,

$$\begin{aligned}
\|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| &\leq \sum_{\ell=2}^{L} \|\delta_\ell(W_2)\|\|h_{\ell-1}(W_2) - h_{\ell-1}(W_1)\| + \sum_{\ell=1}^{L} \|h_{\ell-1}(W_1)\|\|\delta_\ell(W_2) - \delta_\ell(W_1)\|, \\
&\overset{(2)}{\leq} c_2 \sum_{\ell=2}^{L} \|\delta_\ell(W_2)\|\|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| + \sum_{\ell=1}^{L} \|h_{\ell-1}(W_1)\|\|\delta_\ell(W_2) - \delta_\ell(W_1)\|, \\
&= c_2 \sum_{\ell=2}^{L} \|\delta_\ell(W_2)\|\|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| + \sum_{\ell=1}^{L-1} \|h_{\ell-1}(W_1)\|\|\delta_\ell(W_2) - \delta_\ell(W_1)\|, \\
&\quad + \|h_{L-1}(W_1)\|\|\delta_L(W_2) - \delta_L(W_1)\|.
\end{aligned}$$

We bound these three terms. For the first term, we have the sum is bounded as follows,

$$\begin{aligned}
c_2 \sum_{\ell=2}^{L} \|\delta_\ell(W_2)\|\|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| &\overset{\text{Lemma C.12}}{\leq} c_{max}^L \sum_{\ell=2}^{L} \|W_2\|^{L-\ell} \|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\|, \\
&\overset{\text{Lemma C.10}}{\leq} c_{max}^{2L} \sum_{\ell=2}^{L} \|W_2\|^{L-\ell} c_x d_{max}^{1/2} \|W_2 - W_1\| \sum_{k=0}^{\ell-2} \sum_{j=0}^{\ell-k-2} \|W_1\|^j \|W_2\|^k, \\
&= c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{\ell=2}^{L} \sum_{k=0}^{\ell-2} \|W_2\|^{L-\ell+k} \sum_{j=0}^{\ell-k-2} \|W_1\|^j, \\
&= c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{\ell=2}^{L} \sum_{k=L-\ell}^{L-2} \|W_2\|^k \sum_{j=0}^{L-k-2} \|W_1\|^j, \\
&\leq c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| L \sum_{k=0}^{L-2} \|W_2\|^k \sum_{j=0}^{L-k-2} \|W_1\|^j.
\end{aligned}$$

We can bound the third term as

$$\begin{aligned}
\|h_{L-1}(W_1)\|\|\delta_L(W_2)-\delta_L(W_1)\| &\overset{\text{Lemma C.11}}{\leq} c_x d_{max}^{1/2} c_{max}^{L}\Big(\sum_{m=0}^{L-1}\|W_1\|^m\Big)\|\delta_L(W_2)-\delta_L(W_1)\| \\
&\overset{\text{Lemma C.13}}{\leq} c_x^2 d_{max} c_{max}^{2L}\|W_2-W_1\|\Big(\sum_{m=0}^{L-1}\|W_1\|^m\Big)\sum_{k=0}^{L-1}\sum_{j=0}^{L-k-1}\|W_1\|^j\|W_2\|^k, \\
&= c_x^2 d_{max} c_{max}^{2L}\|W_2-W_1\|\sum_{k=0}^{L-1}\|W_2\|^k\sum_{m=0}^{L-1}\sum_{j=0}^{L-k-1}\|W_1\|^{j+m}, \\
&= c_x^2 d_{max} c_{max}^{2L}\|W_2-W_1\|\sum_{k=0}^{L-1}\|W_2\|^k\sum_{m=0}^{L-1}\sum_{j=m}^{L-k-1+m}\|W_1\|^{j}, \\
&\leq L c_x^2 d_{max} c_{max}^{2L}\|W_2-W_1\|\sum_{k=0}^{L-1}\|W_2\|^k\sum_{j=0}^{2L-2-k}\|W_1\|^j.
\end{aligned}$$

For the second term, we have the sum is bounded as

$$\begin{aligned}
\sum_{\ell=1}^{L-1}\|h_{\ell-1}(W_1)\|\|\delta_\ell(W_2)-\delta_\ell(W_1)\| &\overset{\text{Lemma C.11}}{\leq} c_{max}^{L}\sum_{\ell=1}^{L-1} c_x d_{max}^{1/2}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\|\delta_\ell(W_2)-\delta_\ell(W_1)\| \\
&\overset{\text{Lemma C.13}}{\leq} c_{max}^{L}\sum_{\ell=1}^{L-1} c_x d_{max}^{1/2}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\Bigg(c_x d_{max}^{1/2} c_{max}^{2L}\|W_2-W_1\|\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^j \\
&\quad + \sum_{k=0}^{L-\ell-1}\|W_2\|^k\|W_1\|^{L-k-\ell-1}\|W_2-W_1\|\Bigg) \\
&= c_x^2 d_{max} c_{max}^{3L}\|W_2-W_1\|\sum_{\ell=1}^{L-1}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\Bigg(\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^j \\
&\quad + \sum_{k=0}^{L-\ell-1}\|W_2\|^k\|W_1\|^{L-k-\ell-1}\Bigg) \\
&= c_x^2 d_{max} c_{max}^{3L}\|W_2-W_1\|\Bigg[\sum_{\ell=1}^{L-1}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\Bigg(\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^j\Bigg) \\
&\quad + \sum_{\ell=1}^{L-1}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\sum_{k=0}^{L-\ell-1}\|W_2\|^k\|W_1\|^{L-k-\ell-1}\Bigg]
\end{aligned}$$

We have the quintuple sum can be bounded and simplified as follows

$$\begin{aligned}
\sum_{\ell=1}^{L-1}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\Bigg(\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^j\Bigg) &= \sum_{\ell=1}^{L-1}\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{m=0}^{\ell-1}\sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^{m+j},\\
&= \sum_{\ell=1}^{L-1}\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{m=0}^{\ell-1}\sum_{j=L-k+m}^{L-i+k-\ell-1+m}\|W_1\|^j,\\
&\le \sum_{\ell=1}^{L-1}\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{m=0}^{\ell-1}\sum_{j=L-k}^{L-i+k-2}\|W_1\|^j,\\
&\le L\sum_{\ell=1}^{L-1}\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=L-k}^{L-i+k-2}\|W_1\|^j,\\
&\le L\sum_{\ell=1}^{L-1}\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i\sum_{j=0}^{2L-i-2}\|W_1\|^j,\\
&\le L^2\sum_{\ell=1}^{L-1}\sum_{i=0}^{2L-1-\ell}\|W_2\|^i\sum_{j=0}^{2L-i-2}\|W_1\|^j,\\
&\le L^3\sum_{i=0}^{2L-2}\|W_2\|^i\sum_{j=0}^{2L-i-2}\|W_1\|^j.
\end{aligned}$$

And the triple sum can be bounded as

$$\begin{aligned}
\sum_{\ell=1}^{L-1}\Big(\sum_{m=0}^{\ell-1}\|W_1\|^m\Big)\sum_{k=0}^{L-\ell-1}\|W_2\|^k\|W_1\|^{L-k-\ell-1} &= \sum_{\ell=1}^{L-1}\sum_{k=0}^{L-\ell-1}\|W_2\|^k\sum_{m=0}^{\ell-1}\|W_1\|^{L-k-\ell-1+m}\\
&= \sum_{\ell=1}^{L-1}\sum_{k=0}^{L-\ell-1}\|W_2\|^k\sum_{m=L-k-\ell-1}^{L-k-2}\|W_1\|^m\\
&\le \sum_{\ell=1}^{L-1}\sum_{k=0}^{L-\ell-1}\|W_2\|^k\sum_{m=0}^{L-k-2}\|W_1\|^m\\
&\le L\sum_{k=0}^{L-2}\|W_2\|^k\sum_{m=0}^{L-k-2}\|W_1\|^m.
\end{aligned}$$

Putting it all together and using the fact that $L \ge 1$, $d_{max} \ge 1$, $c_x \ge 1$, results in

$$\begin{aligned}
&\|\nabla_W F(x_i;W_2) - \nabla_W F(x_i;W_1)\|\\
&\le L^3 c_x^2 d_{max} c_{max}^{3L}\|W_2 - W_1\|\Bigg(\sum_{k=0}^{L-2}\|W_2\|^k\sum_{j=0}^{L-k-2}\|W_1\|^j\\
&\quad + \sum_{k=0}^{L-1}\|W_2\|^k\sum_{j=0}^{2L-2-k}\|W_1\|^j + \sum_{i=0}^{2L-2}\|W_2\|^i\sum_{j=0}^{2L-i-2}\|W_1\|^j + \sum_{k=0}^{L-2}\|W_2\|^k\sum_{m=0}^{L-k-2}\|W_1\|^m\Bigg)\\
&\le 4c_x^2 d_{max} L^3 c_{max}^{3L}\|W_2 - W_1\|\sum_{k=0}^{2L-2}\|W_2\|^k\sum_{j=0}^{2L-k-2}\|W_1\|^j,\\
&= 4c_x^2 d_{max} L^3 c_{max}^{3L}\|W_2 - W_1\|\sum_{j=0}^{2L-2}\|W_1\|^j\sum_{k=0}^{2L-j-2}\|W_2\|^k.
\end{aligned}$$

where in the last step we exchange the summations.

□

C.6 Proof of Lemma 4.5

Proof. We decompose the change in gradient $\|\nabla\mathcal{L}(W_2) - \nabla\mathcal{L}(W_1)\|$ in terms of $|F(x_i; W_2) - F(x_i; W_1)|$ and $\|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\|$ via Lemma 4.1. In order to leverage Lemma 4.1, we require the following bound on the gradient norm.

$$\begin{aligned}
\|\nabla_W F(x_i; W)\| &\leq \sum_{\ell=1}^{L} \|\delta_\ell(W) h_{\ell-1}(W)^T\|, \\
&\leq \sum_{\ell=1}^{L} \|\delta_\ell(W)\| \|h_{\ell-1}(W)\|, \\
&\overset{\text{Lemmas C.12 and C.11}}{\leq} c_{max}^L \sum_{\ell=1}^{L} \|W\|^{L-\ell} c_x d_{max}^{1/2} \sum_{k=0}^{\ell-1} \|W\|^k, \\
&= c_x d_{max}^{1/2} c_{max}^L \sum_{\ell=1}^{L} \sum_{k=0}^{\ell-1} \|W\|^{L-\ell+k}, \\
&= c_x d_{max}^{1/2} c_{max}^L \sum_{\ell=1}^{L} \sum_{k=L-\ell}^{L-1} \|W\|^k.
\end{aligned}$$

So we have the bound

$$\|\nabla_W F(x_i; W)\| \leq c_x d_{max}^{1/2} c_{max}^L \sum_{\ell=1}^{L} \sum_{k=0}^{L-1} \|W\|^k = c_x d_{max}^{1/2} c_{max}^L L \sum_{k=0}^{L-1} \|W\|^k. \tag{13}$$

Then we can use Lemmas 4.3 and 4.4 to achieve an upper bound in terms of $\|W_2 - W_1\|$, $\|W_1\|$, and $\|W_2\|$.

$$\begin{aligned}
\|\nabla\mathcal{L}(W_2) - \nabla\mathcal{L}(W_1)\| &\overset{\text{Lemma 4.1}}{\leq} \frac{1}{n} \sum_{i=1}^{n} c_J |F(x_i; W_2) - F(x_i; W_1)| \|\nabla_W F(x_i; W_2)\| \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} |J_i'(F(x_i; W_1))| \|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| \\
&\overset{\text{Lemma 4.3 and (13)}}{\leq} c_J c_x d_{max}^{1/2} c_{max}^L \|W_2 - W_1\| \sum_{j=0}^{L-1} \sum_{k=0}^{L-j-1} \|W_1\|^j \|W_2\|^k \cdot c_x d_{max}^{1/2} c_{max}^L L \sum_{q=0}^{L-1} \|W_2\|^q \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} |J_i'(F(x_i; W_1))| \|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| \\
&= c_J c_x^2 d_{max} c_{max}^{2L} L \|W_2 - W_1\| \sum_{j=0}^{L-1} \|W_1\|^j \sum_{k=0}^{L-j-1} \sum_{q=k}^{L-1+k} \|W_2\|^q \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} |J_i'(F(x_i; W_1))| \|\nabla_W F(x_i; W_2) - \nabla_W F(x_i; W_1)\| \\
&\overset{\text{Lemma 4.4}}{\leq} c_J c_x^2 d_{max} c_{max}^{2L} L \|W_2 - W_1\| \sum_{j=0}^{L-1} \|W_1\|^j \sum_{k=0}^{L-j-1} \sum_{q=0}^{2L-2-j} \|W_2\|^q \\
&\quad + 4 c_x^2 d_{max} c_{max}^{3L} L^3 \|W_2 - W_1\| \sum_{j=0}^{2L-2} \|W_1\|^j \sum_{k=0}^{2L-j-2} \|W_2\|^k \cdot \frac{1}{n} \sum_{i=1}^{n} |J_i'(F(x_i; W_1))|.
\end{aligned}$$

From Lipschitz smoothness of J, we have for all W (See Lemma 2.28 in Garrigos & Gower (2024)),

$$|J'(F(x_i; W))|^2 \leq 2 c_J J_i(F(x_i; W)), \tag{14}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|J_i'(F(x_i;W))| \leq& \frac{1}{n}\sum_{i=1}^{n}(2c_J J_i(F(x_i;W)))^{1/2}\\ \leq& (\frac{1}{n}\sum_{i=1}^{n}2c_J J_i(F(x_i;W)))^{1/2}\\ \leq& (2c_J\mathcal{L}(W))^{1/2},\end{aligned}$$

where we use Jensen's inequality and the concavity of the square-root function. So we have

$$\frac{1}{n}\sum_{i=1}^{n}|J_i'(F(x_i;W))| \leq (2c_J\mathcal{L}(W))^{1/2}. \tag{15}$$

Plugging this into the expression, we have

$$\begin{aligned}\|\nabla\mathcal{L}(W_2)-\nabla\mathcal{L}(W_1)\| \overset{(15)}{\leq}& c_J c_x^2 d_{max} c_{max}^{2L} L^2\|W_2-W_1\|\sum_{j=0}^{L-1}\|W_1\|^j\sum_{k=0}^{2L-2-j}\|W_2\|^k\\ &+4(2c_J\mathcal{L}(W_1))^{1/2}c_x^2 d_{max}c_{max}^{3L}L^3\|W_2-W_1\|\sum_{j=0}^{2L-2}\|W_1\|^j\sum_{k=0}^{2L-j-2}\|W_2\|^k,\\ \overset{c_J,c_{max}\geq 1}{\leq}& 8c_J c_x^2 d_{max}c_{max}^{3L}L^3\|W_2-W_1\|\\ &\cdot\left[\sum_{j=0}^{L-1}\|W_1\|^j\sum_{k=0}^{2L-2-j}\|W_2\|^k+\mathcal{L}(W_1)^{1/2}\sum_{j=0}^{2L-2}\|W_1\|^j\sum_{k=0}^{2L-j-2}\|W_2\|^k\right]\\ \leq& 8c_J c_x^2 d_{max}c_{max}^{3L}L^3\|W_2-W_1\|(1+\mathcal{L}(W_1)^{1/2})\sum_{j=0}^{2L-2}\|W_1\|^j\sum_{k=0}^{2L-j-2}\|W_2\|^k.\end{aligned}$$

□

C.7 Proof of Lemma 4.6

Proof. We denote $v = W_2 - W_1$, and parametrize $\mathcal{L}$ between W_1 and W_2 such that for $t \in [0,1]$, we have

$$\begin{aligned}\phi(t) =& \mathcal{L}(W_1+tv),\\ \phi'(t) =& \langle\nabla\mathcal{L}(W_1+tv), v\rangle_F.\end{aligned}$$

By the fundamental theorem of calculus, we have

$$\begin{aligned}\mathcal{L}(W_2)-\mathcal{L}(W_1) = \phi(1)-\phi(0) =& \int_0^1\phi'(t)dt,\\ =& \int_0^1\langle\nabla\mathcal{L}(W_1),v\rangle_F\,dt+\int_0^1\langle\nabla\mathcal{L}(W_1+tv)-\nabla\mathcal{L}(W_1),v\rangle_F\,dt,\\ =& \langle\nabla\mathcal{L}(W_1),v\rangle_F+\int_0^1\langle\nabla\mathcal{L}(W_1+tv)-\nabla\mathcal{L}(W_1),v\rangle_F\,dt,\\ \leq& \langle\nabla\mathcal{L}(W_1),v\rangle_F+\int_0^1\|\nabla\mathcal{L}(W_1+tv)-\nabla\mathcal{L}(W_1)\|\|v\|dt,\end{aligned}$$

where in the last step we use the Cauchy-Schwarz inequality and the fact that integrals preserve inequalities. For the integral, we have by Lemma 4.5,

$$\int_0^1\|\nabla\mathcal{L}(W_1+tv)-\nabla\mathcal{L}(W_1)\|\|v\|dt \leq 8c_J c_x^2 d_{max}c_{max}^{3L}L^3(1+\mathcal{L}(W_1)^{1/2})\|v\|^2\sum_{j=0}^{2L-2}\|W_1\|^j\int_0^1 t\sum_{i=0}^{2L-j-2}\|W_1+tv\|^i dt,$$

where we pull out any terms not dependent on t from the integral. For any nonnegative integer N, we have by the binomial expansion and reindexing

$$\begin{aligned}
\int_0^1 t\sum_{i=0}^{N}\|W_1+tv\|^i\,dt \le& \int_0^1 t\sum_{i=0}^{N}(\|W_1\|+\|tv\|)^i\,dt\\
=&\int_0^1 t\sum_{i=0}^{N}\sum_{k=0}^{i}\binom{i}{k}\|W_1\|^{i-k}t^k\|v\|^k\,dt\\
=&\int_0^1 t\sum_{k=0}^{N}\sum_{i=k}^{N}\binom{i}{k}\|W_1\|^{i-k}t^k\|v\|^k\,dt\\
=&\sum_{k=0}^{N}\sum_{i=k}^{N}\binom{i}{k}\|W_1\|^{i-k}\|v\|^k\int_0^1 t^{k+1}\,dt\\
=&\sum_{k=0}^{N}\sum_{i=k}^{N}\binom{i}{k}\|W_1\|^{i-k}\|v\|^k\frac{1}{k+2}\\
\le& 2^N\sum_{k=0}^{N}\frac{1}{k+2}\|v\|^k\sum_{i=k}^{N}\|W_1\|^{i-k}=2^N\sum_{k=0}^{N}\frac{1}{k+2}\|v\|^k\sum_{i=0}^{N-k}\|W_1\|^i,
\end{aligned}$$

where in the last inequality we use the fact that $\binom{i}{k}\le 2^i\le 2^N$. Therefore we have

$$\begin{aligned}
\sum_{j=0}^{2L-2}\|W_1\|^j\int_0^1 t\sum_{i=0}^{2L-j-2}\|W_1+tv\|^i dt \le& \sum_{j=0}^{2L-2}2^{2L-j-2}\sum_{k=0}^{2L-j-2}\frac{1}{k+2}\|v\|^k\sum_{i=0}^{2L-j-2-k}\|W_1\|^{i+j}\\
=&\sum_{j=0}^{2L-2}2^{2L-j-2}\sum_{k=0}^{2L-j-2}\frac{1}{k+2}\|v\|^k\sum_{i=j}^{2L-2-k}\|W_1\|^i\\
\le& 2^{2L-2}\sum_{j=0}^{2L-2}\sum_{i=j}^{2L-2}\|W_1\|^i\sum_{k=0}^{2L-j-2}\frac{1}{k+2}\|v\|^k\\
\le& 2^{2L-2}(2L-1)\sum_{i=0}^{2L-2}\|W_1\|^i\sum_{k=0}^{2L-2}\frac{1}{k+2}\|v\|^k,
\end{aligned}$$

and the integral term can be bounded as follows,

$$\begin{aligned}
&\int_0^1\|\nabla\mathcal{L}(W_1+tv)-\nabla\mathcal{L}(W_1)\|\|v\|dt\\
&\le 8c_Jc_x^2d_{max}c_{max}^{3L}L^3(1+\mathcal{L}(W_1)^{1/2})\|v\|^2\cdot 2^{2L-2}(2L-1)\sum_{i=0}^{2L-2}\|W_1\|^i\sum_{k=0}^{2L-2}\frac{1}{k+2}\|v\|^k\\
&\le 2^{2L+2}c_Jc_x^2d_{max}c_{max}^{3L}L^4(1+\mathcal{L}(W_1)^{1/2})\sum_{i=0}^{2L-2}\|W_1\|^i\sum_{k=0}^{2L-2}\frac{1}{k+2}\|v\|^{k+2}.
\end{aligned}$$

Let $C=2^{2L+2}c_Jc_x^2d_{max}c_{max}^{3L}L^4$. Then putting it all together gives

$$\begin{aligned}
\mathcal{L}(W_2)-\mathcal{L}(W_1)\le&\langle\nabla\mathcal{L}(W_1),W_2-W_1\rangle_F\\
&+C\Big[(1+\mathcal{L}(W_1)^{1/2})\sum_{i=0}^{2L-2}\|W_1\|^i\Big]\sum_{k=0}^{2L-2}\frac{1}{k+2}\|W_2-W_1\|^{k+2}.
\end{aligned}$$

□

C.8 Proof of Lemma 4.7

Applying the descent lemma to W_{t+1} and W_t results in sums of powers of η_t. In the following, we set η_t small enough by considering the coefficients of the η_t^2 terms. We then demonstrate that the

terms corresponding to higher powers of η_t are also controlled to be small, resulting in descent in function value.

For $W_{t+1} = W_t - \eta_t \nabla \mathcal{L}(W_t)$, we have by Lemma 4.6

$$\begin{aligned}\mathcal{L}(W_{t+1}) - \mathcal{L}(W_t) \leq & \langle \nabla \mathcal{L}(W_t), W_{t+1} - W_t \rangle_F + C(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i \sum_{k=0}^{2L-2} \frac{1}{k+2} \|W_{t+1} - W_t\|^{k+2} \\ = & -\eta_t \|\nabla \mathcal{L}(W_t)\|^2 + C(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i \sum_{k=0}^{2L-2} \frac{1}{k+2} \|\nabla \mathcal{L}(W_t)\|^{k+2} \eta_t^{k+2}.\end{aligned}$$

We observe the right hand side consists of $2L-1$ nonnegative higher-order terms of the form $\eta_t^\gamma \|\nabla \mathcal{L}(W_t)\|^\gamma$, for $\gamma = 2, ..., 2L$. We set

$$\eta_t = \frac{1}{2LC(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i} = \frac{1}{\rho(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i}$$

where $\rho = 2LC$. We want to demonstrate that η_t is small enough such that each term is less than or equal to $\frac{\eta_t}{2L} \|\nabla \mathcal{L}(W_t)\|^2$. The $\gamma = 2$ case of $\eta_t^2 \|\nabla \mathcal{L}(W_t)\|^2$ is trivial. We have

$$\begin{aligned}\eta_t^2 \|\nabla \mathcal{L}(W_t)\|^2 \frac{C}{2} (1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i \leq & \eta_t \cdot \frac{\frac{C}{2}(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i}{2LC(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i} \|\nabla \mathcal{L}(W_t)\|^2 \\ \leq & \frac{\eta_t}{2L} \|\nabla \mathcal{L}(W_t)\|^2.\end{aligned}$$

Now for $k = 1, ..., 2L-2$, where $k+2 = \gamma$, we want to show the following statement,

$$\eta_t^{k+2} \|\nabla \mathcal{L}(W_t)\|^{k+2} \frac{C}{k+2} (1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i \leq \frac{\eta_t}{2L} \|\nabla \mathcal{L}(W_t)\|^2.$$

We have the gradient norm is bounded as

$$\begin{aligned}\|\nabla \mathcal{L}(W_t)\| \leq & \frac{1}{n} \sum_{i=1}^n |J_i'(F(x_i; W))| \|\nabla_W F(x_i; W)\| \\ \overset{(13)}{\leq} & \frac{1}{n} \sum_{i=1}^n |J_i'(F(x_i; W))| c_x d_{max}^{1/2} c_{max}^L L \sum_{k=0}^{L-1} \|W\|^k \\ \overset{(15)}{\leq} & (2c_J \mathcal{L}(W))^{1/2} \cdot c_x d_{max}^{1/2} c_{max}^L L \sum_{k=0}^{L-1} \|W\|^k,\end{aligned}$$

and we therefore have the bound

$$\eta_t \|\nabla \mathcal{L}(W_t)\| \overset{(13)}{\leq} \frac{(2c_J \mathcal{L}(W_t))^{1/2} c_x d_{max}^{1/2} c_{max}^L L \sum_{k=0}^{L-1} \|W_t\|^k}{2LC(1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i} \leq \frac{1}{2^{2L+2} c_J^{1/2} c_x d_{max}^{1/2} c_{max}^{2L} L^4} \leq 1, \tag{16}$$

so we can bound powers of this term,

$$\begin{aligned}& \eta_t^{k+2} \|\nabla \mathcal{L}(W_t)\|^{k+2} \frac{C}{k+2} (1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i \\ & = \Big(\eta_t \|\nabla \mathcal{L}(W_t)\|^2\Big) \Big(\eta_t \frac{C}{k+2} (1 + \mathcal{L}(W_t)^{1/2}) \sum_{i=0}^{2L-2} \|W_t\|^i\Big) \eta_t^k \|\nabla \mathcal{L}(W_t)\|^k \\ & \leq \Big(\eta_t \|\nabla \mathcal{L}(W_t)\|^2\Big) \frac{1}{2L} \eta_t^k \|\nabla \mathcal{L}(W_t)\|^k \\ & \overset{(16)}{\leq} \Big(\eta_t \|\nabla \mathcal{L}(W_t)\|^2\Big) \frac{1}{2L} \leq \frac{\eta_t}{2L} \|\nabla \mathcal{L}(W_t)\|^2.\end{aligned}$$

So we have shown that each of the $2L-1$ terms is less than or equal to $\frac{\eta_t}{2L}\|\nabla\mathcal{L}(W_t)\|^2$. We have

$$\begin{aligned}\mathcal{L}(W_{t+1}) - \mathcal{L}(W_t) \leq &- \eta_t\|\nabla\mathcal{L}(W_t)\|^2 + \eta_t\frac{2L-1}{2L}\|\nabla\mathcal{L}(W_t)\|^2 \\ = &- \frac{\eta_t}{2L}\|\nabla\mathcal{L}(W_t)\|^2.\end{aligned}$$

Rearranging and summing on both sides yields

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{\eta_t}{2L}\|\nabla\mathcal{L}(W_t)\|^2 \leq &\sum_{t=0}^{T-1}\mathcal{L}(W_t) - \mathcal{L}(W_{t+1}) \\ \sum_{t=0}^{T-1}\eta_t\|\nabla\mathcal{L}(W_t)\|^2 \leq &2L(\mathcal{L}(W_0) - \mathcal{L}(W_T)).\end{aligned}$$

Since $\mathcal{L}(W_T) \geq 0$, we have the following bound, which we will leverage multiple times in the proof:

$$\sum_{t=0}^{T-1}\eta_t\|\nabla\mathcal{L}(W_t)\|^2 \leq 2L\mathcal{L}(W_0). \tag{17}$$

We can therefore bound the minimum squared gradient norm as follows.

$$\begin{aligned}\min_{t=0,\dots,T-1}\|\nabla\mathcal{L}(W_t)\|^2\sum_{t=0}^{T-1}\eta_t \leq &2L\mathcal{L}(W_0), \\ \min_{t=0,\dots,T-1}\|\nabla\mathcal{L}(W_t)\|^2 \leq &\frac{2L\mathcal{L}(W_0)}{\sum_{t=0}^{T-1}\eta_t}.\end{aligned}$$

C.9 Proof of Theorem 4.8

To show convergence of the gradient norm to zero, we need to show that the sum of learning rates $\sum_{t=0}^{T-1}\eta_t$ diverges. We achieve this by lower bounding with another series, which we show diverges. First, we note that since $\mathcal{L}(W_t)$ is nonincreasing with t, we have $(\mathcal{L}(W_t))^{1/2} \leq (\mathcal{L}(W_0))^{1/2}$. Moreover, we have for all $0 \leq j \leq N$,

$$\|W\|^j \leq \|W\|^N + 1. \tag{18}$$

So we have for $N = 2L-2$,

$$\sum_{i=0}^{2L-2}\|W_t\|^i \leq (2L-1)(1+\|W_t\|^N).$$

Let $Q_t = \rho(1+\mathcal{L}(W_0)^{1/2})(2L-1)(1+\|W_t\|^N)$. Then we have

$$\eta_t = \frac{1}{\rho(1+\mathcal{L}(W_t)^{1/2})\sum_{i=0}^{2L-2}\|W_t\|^i} \geq \frac{1}{Q_t}$$

To show the divergence of $\sum_{t=0}^{\infty}\eta_t$, it suffices to show the divergence of $\sum_{t=0}^{\infty}\frac{1}{Q_t}$. We therefore want to track the growth of $\|W_t\|^N$ as t increases.

We first observe that for any $0 \leq j \leq N \leq 2L-2$, that

$$\|W_t\|^j\eta_t^{\frac{j}{N}} \leq \|W_t\|^j\Big(\frac{1}{\rho\|W_t\|^N}\Big)^{\frac{j}{N}} \leq 1, \tag{19}$$

where we also use that $\rho > 1$. We can therefore bound for any $N \leq 2L-2$,

$$\begin{aligned}
\|W_{t+1}\|^N =& \|W_t - \eta_t \nabla \mathcal{L}(W_t)\|^N \\
\leq& (\|W_t\| + \eta_t \|\nabla \mathcal{L}(W_t)\|)^N \\
=& \sum_{k=0}^{N} \binom{N}{k} \|W_t\|^{N-k} \eta_t^k \|\nabla \mathcal{L}(W_t)\|^k \\
=& \|W_t\|^N + N\|W_t\|^{N-1}\eta_t \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} \|W_t\|^{N-k} \eta_t^k \|\nabla \mathcal{L}(W_t)\|^k \\
=& \|W_t\|^N + N\|W_t\|^{N-1}\eta_t^{\frac{N-1}{N}} \eta_t^{\frac{1}{N}} \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} \|W_t\|^{N-k} \eta_t^{k/2} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k \\
\overset{\eta_t \leq 1}{\leq}& \|W_t\|^N + N\|W_t\|^{N-1}\eta_t^{\frac{N-1}{N}} \eta_t^{\frac{1}{N}} \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} \|W_t\|^{N-k} \eta_t \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k \\
\overset{(19)}{\leq}& \|W_t\|^N + N\eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k \\
\|W_{t+1}\|^N - \|W_t\|^N \leq& N\eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k
\end{aligned}$$

When we sum from $t = 0$ to $T-1$ on both sides, we achieve a telescoping sum as follows,

$$\begin{aligned}
\sum_{t=0}^{T-1} \|W_{t+1}\|^N - \|W_t\|^N \leq& N \sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{t=0}^{T-1} \sum_{k=2}^{N} \binom{N}{k} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k, \\
\|W_T\|^N \leq& \|W_0\|^N + N \sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{t=0}^{T-1} \sum_{k=2}^{N} \binom{N}{k} \eta_t^{k/2} \|\nabla \mathcal{L}(W_t)\|^k.
\end{aligned}$$

The double sum last term can be bounded via (17) as follows (see also, Lemma C.5)

$$\|W_T\|^N \leq \|W_0\|^N + N \sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| + \sum_{k=2}^{N} \binom{N}{k} (2L\mathcal{L}(W_0))^{k/2}. \tag{20}$$

The crux of this proof is bounding $\sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\|$. When $N = 2$, which is true in the two-layer case, this term can be directly bounded using (17), leading to an $O(\sqrt{T})$ growth rate. However, for $N > 2$, we need to control the cumulative terms more carefully. For $N = 2L-2$, we can decompose the sum as follows:

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t^{1/N} \|\nabla \mathcal{L}(W_t)\| =& \sum_{t=0}^{T-1} \eta_t^{1/N - 1/2} \eta_t^{1/2} \|\nabla \mathcal{L}(W_t)\| = \sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{2N}} \eta_t^{1/2} \|\nabla \mathcal{L}(W_t)\|, \\
\overset{\text{Cauchy-Schwarz}}{\leq}& \Big(\sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{N}} \Big)^{1/2} \Big(\sum_{t=0}^{T-1} \eta_t \|\nabla \mathcal{L}(W_t)\|^2 \Big)^{1/2} \\
\overset{(17)}{\leq}& \Big(\sum_{t=0}^{T-1} \eta_t^{\frac{2-N}{N}} \Big)^{1/2} (2L\mathcal{L}(W_0))^{1/2}.
\end{aligned}$$

We have

$$\begin{aligned}\eta_t^{\frac{2-N}{N}} =&\Big(\frac{1}{\eta_t}\Big)^{\frac{N-2}{N}} = (\rho(1+\mathcal{L}(W_t)^{1/2})\sum_{i=0}^{2L-2}\|W_t\|^i)^{\frac{N-2}{N}}\\ \overset{\text{Lemma 4.7}}{\leq}&(\rho(1+\mathcal{L}(W_0)^{1/2})\sum_{i=0}^{2L-2}\|W_t\|^i)^{\frac{N-2}{N}}\\ \overset{(18)}{\leq}&(\rho(1+\mathcal{L}(W_0)^{1/2})\sum_{i=0}^{2L-2}(1+\|W_t\|^N))^{\frac{N-2}{N}}\\ =&(\rho(1+\mathcal{L}(W_0)^{1/2})(2L-1)(1+\|W_t\|^N))^{\frac{N-2}{N}}\end{aligned}$$

Plugging this back into (20), we have

$$\begin{aligned}\|W_T\|^N \leq&\|W_0\|^N + N\Big(\sum_{t=0}^{T-1}(\rho(1+\mathcal{L}(W_0)^{1/2})(2L-1)(1+\|W_t\|^N))^{\frac{N-2}{N}}\Big)^{1/2}(2L\mathcal{L}(W_0))^{1/2}\\ &+\sum_{k=2}^{N}\binom{N}{k}(2L\mathcal{L}(W_0))^{k/2},\\ =&\|W_0\|^N + N(\rho(1+\mathcal{L}(W_0)^{1/2})(2L-1))^{\frac{N-2}{2N}}\Big(\sum_{t=0}^{T-1}(1+\|W_t\|^N)^{\frac{N-2}{N}}\Big)^{1/2}(2L\mathcal{L}(W_0))^{1/2}\\ &+\sum_{k=2}^{N}\binom{N}{k}(2L\mathcal{L}(W_0))^{k/2}.\end{aligned}$$

We therefore have that $\|W_T\|^N$ is bounded as

$$\|W_T\|^N \leq A\Big(\sum_{t=0}^{T-1}(1+\|W_t\|^N)^{\frac{N-2}{N}}\Big)^{1/2} + B.$$

where $A = N(\rho(1+\mathcal{L}(W_0)^{1/2})(2L-1))^{\frac{N-2}{2N}}(2L\mathcal{L}(W_0))^{1/2}$ and $B = \|W_0\|^N + \sum_{k=2}^{N}\binom{N}{k}(2L\mathcal{L}(W_0))^{k/2}$.

So the growth of $\|W_T\|^N$ depends on the growth of the $t = 0, ..., T-1$ iterates before; for example, as long as those can be shown to grow at most linearly, then $\|W_T\|^N$ will also be linear in T. In particular, we can show the growth is *sublinear* by proving with induction that for all $T > 0$, there exists $\alpha > 0$, $\beta > 0$ such that

$$\|W_T\|^N \leq \alpha T^{\frac{N}{N+2}} + \beta.$$

Let $\beta = B$. For the base case, we have

$$\|W_0\|^N \leq B.$$

For the inductive step, suppose the hypothesis holds for $t \leq T-1$. We have

$$\begin{aligned}
\|W_T\|^N \leq & A\Big(\sum_{t=0}^{T-1}(1+\alpha t^{\frac{N}{N+2}}+\beta)^{\frac{N-2}{N}}\Big)^{1/2}+\beta \\
\overset{\frac{N-2}{N}<1}{\leq} & A\Big(\sum_{t=0}^{T-1}(1+\beta)^{\frac{N-2}{N}}+(\alpha t^{\frac{N}{N+2}})^{\frac{N-2}{N}}\Big)^{1/2}+\beta \\
= & A\Big(\sum_{t=0}^{T-1}((1+\beta)^{\frac{N-2}{N}}+\alpha^{\frac{N-2}{N}}t^{\frac{N-2}{N+2}})\Big)^{1/2}+\beta \\
\leq & A\Big(T(1+\beta)^{\frac{N-2}{N}}+T\cdot\alpha^{\frac{N-2}{N}}T^{\frac{N-2}{N+2}}\Big)^{1/2}+\beta \\
= & A\Big(T(1+\beta)^{\frac{N-2}{N}}+\alpha^{\frac{N-2}{N}}T^{\frac{2N}{N+2}}\Big)^{1/2}+\beta \\
\overset{\frac{2N}{N+2}>1}{\leq} & A\Big(((1+\beta)^{\frac{N-2}{N}}+\alpha^{\frac{N-2}{N}})T^{\frac{2N}{N+2}}\Big)^{1/2}+\beta \\
= & A((1+\beta)^{\frac{N-2}{N}}+\alpha^{\frac{N-2}{N}})^{1/2}T^{\frac{N}{N+2}}+\beta \\
\leq & A((1+\beta)^{\frac{N-2}{2N}}+\alpha^{\frac{N-2}{2N}})T^{\frac{N}{N+2}}+\beta.
\end{aligned}$$

To achieve the proof, we need to show there exists α such that

$$\begin{aligned}
A((1+\beta)^{\frac{N-2}{2N}}+\alpha^{\frac{N-2}{2N}})T^{\frac{N}{N+2}}+\beta \leq & \alpha T^{\frac{N}{N+2}}+\beta \\
A((1+\beta)^{\frac{N-2}{2N}}+\alpha^{\frac{N-2}{2N}}) \leq & \alpha.
\end{aligned}$$

Since the right hand side grows linearly in α while the left hand side is sublinear, growing as $\alpha^{\frac{N-2}{2N}}$, we are guaranteed to find α^* large enough to satisfy this condition. We therefore have that $\sum_{t=0}^{T-1}\frac{1}{Q_t}$ is lower bounded by a p-series that grows as $T^{1/L}$, which also lower bounds $\sum_{t=0}^{T-1}\eta_t$.

$$\sum_{t=0}^{T-1}\eta_t \geq \sum_{t=0}^{T-1}\frac{1}{Q_t} \geq \sum_{t=0}^{T-1}\frac{1}{\rho(1+\mathcal{L}(W_0)^{1/2})(2L-1)(1+\alpha^* t^{\frac{L-1}{L}}+\beta)} = \Theta(T^{1/L}).$$

We therefore have that

$$\min_{t=0,\ldots,T-1}\|\nabla\mathcal{L}(W_t)\|^2 \leq \frac{2L\mathcal{L}(W_0)}{\sum_{t=0}^{T-1}\eta_t} = O(\frac{1}{T^{1/L}}),$$

which concludes the proof.

C.10 Bounds on $z_\ell(W)$, $h_\ell(W)$, and $\delta_\ell(W)$

Lemma C.9. *For $z_\ell(W)$ defined in (7), we have for all $W \in \mathbb{R}^{D_2\times D_1}$,*

$$\|z_\ell(W)\| \leq c_x d_{max}^{1/2} c_{max}^{\ell-1}\sum_{i=1}^{\ell}\|W\|^i.$$

Proof. For the base case $\ell = 1$, we have

$$\|z_1(W)\| = \|E_1[W]x_i\| \leq \|W\|\|x_i\| \leq \|W\|c_x d^{1/2}.$$

For the general case, we have

$$\begin{aligned}
\|z_\ell(W)\| = & \|E_\ell[W]\sigma_{\ell-1}(z_{\ell-1}(W))\|, \\
\leq & \|W\|\|\sigma_{\ell-1}(z_{\ell-1}(W))\|, \\
\overset{\text{Lemma C.1}}{\leq} & \|W\|(c_0 d_{\ell-1}^{1/2}+c_1\|z_{\ell-1}(W)\|), \\
\leq & c_x d_{max}^{1/2} c_{max}^{\ell-1}\sum_{i=1}^{\ell}\|W\|^i.
\end{aligned}$$

□

Lemma C.10. *For all $W_1, W_2 \in \mathbb{R}^{D_2 \times D_1}$ and $\ell = 1, ..., L$, we have*

$$\|z_\ell(W_2) - z_\ell(W_1)\| \leq c_x d_{max}^{1/2} c_{max}^{\ell-1} \|W_2 - W_1\| \sum_{k=0}^{\ell-1} \sum_{j=0}^{\ell-k-1} \|W_1\|^j \|W_2\|^k.$$

Proof. We first show the bound for $\ell = 1$.

$$\begin{aligned}\|z_1(W_2) - z_1(W_1)\| =& \|E_1[W_2]x_i - E_1[W_1]x_i\| \\ \leq& c_x d^{1/2} \|W_2 - W_1\| \leq c_x d_{max}^{1/2} \|W_2 - W_1\|.\end{aligned}$$

Then for general $\ell = 2, ..., L$, we have

$$\begin{aligned}\|z_\ell(W_2) - z_\ell(W_1)\| =& \|E_\ell[W_2]\sigma_{\ell-1}(z_{\ell-1}(W_2)) - E_\ell[W_1]\sigma_{\ell-1}(z_{\ell-1}(W_1))\|, \\ \leq& \|W_2\| \|\sigma_{\ell-1}(z_{\ell-1}(W_2)) - \sigma_{\ell-1}(z_{\ell-1}(W_1))\| + \|\sigma_{\ell-1}(z_{\ell-1}(W_1))\| \|W_2 - W_1\|, \\ \leq& c_2 \|W_2\| \|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| + \|\sigma_{\ell-1}(z_{\ell-1}(W_1))\| \|W_2 - W_1\|, \\ \overset{\text{Lemma C.1}}{\leq}& c_2 \|W_2\| \|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| + (c_0 d_{\ell-1}^{1/2} + c_1 \|z_{\ell-1}(W_1)\|) \|W_2 - W_1\|, \\ \overset{\text{Lemma C.9}}{\leq}& c_2 \|W_2\| \|z_{\ell-1}(W_2) - z_{\ell-1}(W_1)\| + c_x d_{max}^{1/2} c_{max}^{\ell-1} \Big(\sum_{i=0}^{\ell-1} \|W_1\|^i \Big) \|W_2 - W_1\|, \\ \leq& c_x d_{max}^{1/2} c_{max}^{\ell-1} \|W_2 - W_1\| \sum_{k=0}^{\ell-1} \sum_{j=0}^{\ell-k-1} \|W_1\|^j \|W_2\|^k.\end{aligned}$$

□

Lemma C.11. *For $h_\ell(W)$ defined in (7), we have for all $W \in \mathbb{R}^{D_2 \times D_1}$,*

$$\|h_\ell(W)\| \leq c_x d_{max}^{1/2} c_{max}^{\ell} \sum_{k=0}^{\ell} \|W\|^k.$$

Proof. For the base case $\ell = 0$, we have

$$\|h_0(W)\| = \|x_i\| \leq d^{1/2} c_x.$$

For the general case, we have by Lemma C.9,

$$\begin{aligned}\|h_\ell(W)\| =& \|\sigma_\ell(z_\ell(W))\|, \\ \overset{\text{Lemma C.1}}{\leq}& c_0 d_\ell^{1/2} + c_1 \|z_\ell(W)\|, \\ \leq& c_0 d_\ell^{1/2} + c_x d_{max}^{1/2} c_{max}^{\ell} \sum_{i=1}^{\ell} \|W\|^i, \\ \leq& c_x d_{max}^{1/2} c_{max}^{\ell} \sum_{k=0}^{\ell} \|W\|^k,\end{aligned}$$

using the fact that $c_x \geq 1$. □

Lemma C.12. *For $\delta_\ell(W)$ defined in (8), we have for all $W \in \mathbb{R}^{D_2 \times D_1}$,*

$$\|\delta_\ell(W)\| \leq c_{max}^{L-\ell+1} \|W\|^{L-\ell}.$$

Proof. For the base case $\ell = L$, we have the following due to Assumption 3.1 and $d_L = 1$,

$$\|\delta_L(W)\| = \|\sigma_L'(z_L(W))\| \leq c_2.$$

For general ℓ, we have

$$\begin{aligned}\|\delta_\ell(W)\| =&\|\sigma'_\ell(z_\ell(W)) \odot (E_{\ell+1}[W]^T\delta_{\ell+1}(W))\|,\\ \overset{\text{Lemma C.8}}{\leq}\ & c_2\|E_{\ell+1}[W]^T\delta_{\ell+1}(W)\|,\\ \leq& c_2\|W\|\|\delta_{\ell+1}(W)\|,\\ \leq& c_{max}^{L-\ell+1}\|W\|^{L-\ell}.\end{aligned}$$

□

Lemma C.13. *For all $W_1, W_2 \in \mathbb{R}^{D_2\times D_1}$, we have*

$$\|\delta_L(W_2)-\delta_L(W_1)\| \leq \|z_L(W_2)-z_L(W_1)\| \leq c_x d_{max}^{1/2} c_{max}^L \|W_2-W_1\| \sum_{k=0}^{L-1}\sum_{j=0}^{L-k-1}\|W_1\|^j\|W_2\|^k.$$

and for $\ell = 1, .., L-1$, we have

$$\|\delta_\ell(W_2)-\delta_\ell(W_1)\| \leq c_x d_{max}^{1/2} c_{max}^{2L} \|W_2-W_1\| \Big[\sum_{k=\ell}^{L}\sum_{i=k-\ell}^{2k-1-\ell}\|W_2\|^i \sum_{j=L-k}^{L-i+k-\ell-1}\|W_1\|^j + \sum_{k=0}^{L-\ell-1}\|W_2\|^k\|W_1\|^{L-k-\ell-1}\Big].$$

Proof. For the base case $\ell = L$, we have

$$\begin{aligned}\|\delta_L(W_2)-\delta_L(W_1)\| =&\|\sigma'_L(z_L(W_2)) - \sigma'_L(z_L(W_1))\|\\ \overset{(2)}{\leq}& c_3\|z_L(W_2)-z_L(W_1)\|,\\ \overset{\text{Lemma C.10}}{\leq}& c_x d_{max}^{1/2} c_{max}^L \|W_2-W_1\| \sum_{k=0}^{L-1}\sum_{j=0}^{L-k-1}\|W_1\|^j\|W_2\|^k.\end{aligned}$$

For $\ell \leq L-1$, we have

$$\begin{aligned}\|\delta_\ell(W_2)-\delta_\ell(W_1)\| =&\|\sigma'_\ell(z_\ell(W_2)) \odot (E_{\ell+1}[W_2]^T\delta_{\ell+1}(W_2)) - \sigma'_\ell(z_\ell(W_1)) \odot (E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1))\|\\ \leq&\|\sigma'_\ell(z_\ell(W_2)) \odot (E_{\ell+1}[W_2]^T\delta_{\ell+1}(W_2) - E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1))\|\\ &+ \|\sigma'_\ell(z_\ell(W_2)) - \sigma'_\ell(z_\ell(W_1))\|\|E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1)\|\\ \leq& c_2\|E_{\ell+1}[W_2]^T\delta_{\ell+1}(W_2) - E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1)\|\\ &+ \|\sigma'_\ell(z_\ell(W_2)) - \sigma'_\ell(z_\ell(W_1))\|\|E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1)\|\\ \overset{(2)}{\leq}& c_2\|E_{\ell+1}[W_2]^T\delta_{\ell+1}(W_2) - E_{\ell+1}[W_1]^T\delta_{\ell+1}(W_1)\|\\ &+ c_3\|z_\ell(W_2)-z_\ell(W_1)\|\|W_1\|\|\delta_{\ell+1}(W_1)\|\\ \leq& c_2\|\delta_{\ell+1}(W_1)\|\|W_2-W_1\| + c_2\|W_2\|\|\delta_{\ell+1}(W_2)-\delta_{\ell+1}(W_1)\|\\ &+ c_3\|z_\ell(W_2)-z_\ell(W_1)\|\|W_1\|\|\delta_{\ell+1}(W_1)\|\\ \overset{\text{Lemma C.12}}{\leq}& c_{max}^{L-\ell+1}\|W_1\|^{L-\ell-1}\|W_2-W_1\| + c_2\|W_2\|\|\delta_{\ell+1}(W_2)-\delta_{\ell+1}(W_1)\|\\ &+ c_{max}^{L-\ell+1}\|z_\ell(W_2)-z_\ell(W_1)\|\|W_1\|\|W_1\|^{L-\ell-1}\\ =& c_{max}^{L-\ell+1}\|W_1\|^{L-\ell-1}\|W_2-W_1\| + c_2\|W_2\|\|\delta_{\ell+1}(W_2)-\delta_{\ell+1}(W_1)\|\\ &+ c_{max}^{L-\ell+1}\|z_\ell(W_2)-z_\ell(W_1)\|\|W_1\|^{L-\ell}\end{aligned}$$

Solving the recursive relationship leads to

$$\begin{aligned}
\|\delta_\ell(W_2) - \delta_\ell(W_1)\| \leq & c_{max}^L \sum_{k=\ell}^{L} \|W_2\|^{k-\ell} \|z_k(W_2) - z_k(W_1)\| \|W_1\|^{L-k} + c_{max}^L \sum_{k=\ell}^{L-1} \|W_2\|^{k-\ell} \|W_1\|^{L-k-1} \|W_2 - W_1\| \\
\overset{\text{Lemma C.10}}{\leq} & c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{k=\ell}^{L} \|W_2\|^{k-\ell} \Big(\sum_{i=0}^{k-1} \sum_{j=0}^{k-i-1} \|W_1\|^j \|W_2\|^i\Big) \|W_1\|^{L-k} \\
& + c_{max}^L \sum_{k=\ell}^{L-1} \|W_2\|^{k-\ell} \|W_1\|^{L-k-1} \|W_2 - W_1\| \\
= & c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{k=\ell}^{L} \sum_{i=0}^{k-1} \sum_{j=0}^{k-i-1} \|W_1\|^{L-k+j} \|W_2\|^{k-\ell+i} \\
& + c_{max}^L \sum_{k=\ell}^{L-1} \|W_2\|^{k-\ell} \|W_1\|^{L-k-1} \|W_2 - W_1\| \\
= & c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{k=\ell}^{L} \sum_{i=k-\ell}^{2k-1-\ell} \|W_2\|^i \sum_{j=L-k}^{L-i+k-\ell-1} \|W_1\|^j \\
& + c_{max}^L \sum_{k=\ell}^{L-1} \|W_2\|^{k-\ell} \|W_1\|^{L-k-1} \|W_2 - W_1\| \\
= & c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \sum_{k=\ell}^{L} \sum_{i=k-\ell}^{2k-1-\ell} \|W_2\|^i \sum_{j=L-k}^{L-i+k-\ell-1} \|W_1\|^j \\
& + c_{max}^L \sum_{k=0}^{L-\ell-1} \|W_2\|^k \|W_1\|^{L-k-\ell-1} \|W_2 - W_1\| \\
\leq & c_x d_{max}^{1/2} c_{max}^{2L} \|W_2 - W_1\| \Big[\sum_{k=\ell}^{L} \sum_{i=k-\ell}^{2k-1-\ell} \|W_2\|^i \sum_{j=L-k}^{L-i+k-\ell-1} \|W_1\|^j + \sum_{k=0}^{L-\ell-1} \|W_2\|^k \|W_1\|^{L-k-\ell-1} \Big].
\end{aligned}$$

□

C.11 Lipschitz properties of common activation and loss functions

In the following, we demonstrate that the following popular activation and loss functions satisfy the assumptions of this work.

Linear activation function. The linear activation function is defined as

$$\sigma(x) = x.$$

This is linearly bounded with $c_0 = 0$ and $c_1 = 1$. Since $\sigma'(x) = 1$ and $\sigma''(x) = 0$, it is Lipschitz continuous and Lipschitz smooth.

Tanh activation function. The tanh activation function is defined as

$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

It is uniformly bounded $|\sigma(x)| \leq 1$ so it is also linearly bounded. We also have that $|\sigma'(x)| \leq 1$ and $|\sigma''(x)| \leq \frac{4}{3\sqrt{3}}$ so it is Lipschitz continuous and Lipschitz smooth.

Sigmoid activation function. The sigmoid activation function is defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

It is is uniformly bounded $|\sigma(x)| \leq 1$ so it is also linearly bounded. We also have that $|\sigma'(x)| \leq \frac{1}{4}$ and $|\sigma''(x)| \leq \frac{1}{6\sqrt{3}}$, so it is Lipschitz continuous and Lipschitz smooth.

Softplus activation function. The softplus activation function is defined as

$$\sigma(x) = \ln(1 + e^x).$$

We have that $|\sigma(x)| \leq \ln(2) + |x|$, so it is linearly bounded. We also have $|\sigma'(x)| \leq 1$ and $|\sigma''(x)| \leq \frac{1}{4}$, so it is Lipschitz continuous and Lipschitz smooth.

Gaussian Error Linear Unit (GELU) activation function. The GELU function (Hendrycks & Gimpel, 2016) weighs an input by its value and cumulative probability under a standard normal distribution, resulting in a smoothed alternative to ReLU activation functions. The exact definition using the standard normal cumulative distribution function $\Phi(x)$ is as follows,

$$\sigma(x) = x\Phi(x) = \frac{1}{2}x\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right).$$

We have that $|\sigma(x)| \leq |x|$, so it is linearly bounded. We also have

$$|\sigma'(x)| = |\frac{1}{2}(\text{erf}(\frac{x}{\sqrt{2}}) + 1) + \frac{xe^{-x^2/2}}{\sqrt{2\pi}}| \leq \frac{1}{2}(\text{erf}(1) + 1 + \frac{2}{e\sqrt{\pi}}) \approx 1.1289,$$

and

$$|\sigma''(x)| = |-\frac{e^{-x^2/2}(x^2 - 2)}{\sqrt{2\pi}}| \leq \sqrt{2/\pi},$$

so it is Lipschitz continuous and Lipschitz smooth.

Sigmoid Linear Unit (SiLU) activation function. The SiLU function (Elfwing et al., 2018), also known as Swish, is another smooth alternative to ReLU which is defined as the product of x and the sigmoid function as follows,

$$\sigma(x) = x\big(\frac{1}{1 + e^{-x}}\big).$$

We have that $|\sigma(x)| \leq x$, so it is linearly bounded. We also have that $|\sigma'(x)|$ achieves a maximum value of approximately 1.0998 and $|\sigma''(x)| \leq \frac{1}{4}$. So this function is Lipschitz continuous and Lipschitz smooth.

Huber loss. The Huber loss function is a differentiable function that combines the properties of mean squared error and mean absolute error. For parameter $\tau > 0$, the Huber loss function is defined as

$$J_i(u) = \begin{cases} \frac{1}{2}(u - y_i)^2 & \text{if } |u - y_i| \leq \tau \\ \tau|u - y_i| - \frac{1}{2}\tau^2 & \text{otherwise.} \end{cases}$$

The derivative of J_i is the following continuous function

$$J_i'(u) = \begin{cases} u - y_i & \text{if } |u - y_i| \leq \tau \\ \tau sgn(u - y_i) & \text{otherwise.} \end{cases}$$

We note that the second derivative is uniformly bounded everywhere such that $|J_i''(u)| \leq 1$ for all $u \in \mathbb{R}$. So Huber loss satisfies Assumption 3.2 with $c_J = 1$.

Binary cross-entropy loss. We have the following extended derivation from the main body.

$$\begin{aligned} J_i(u) =& -\big[y_i \log(\frac{1}{1 + e^{-u}}) + (1 - y_i)\log(1 - \frac{1}{1 + e^{-u}})\big], \\ =& y_i \log(1 + e^{-u}) - (1 - y_i)\log(\frac{e^{-u}}{1 + e^{-u}}) \\ =& y_i \log(1 + e^{-u}) + (y_i - 1)[\log(e^{-u}) - \log(1 + e^{-u})] \\ =& (y_i - 1)(-u) + \log(1 + e^{-u}) \\ =& -uy_i + \log(e^u) + \log(1 + e^{-u}) \\ =& -uy_i + \log(1 + e^u). \end{aligned}$$

D EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

Code for all experiments is open-sources at the anonymous Github repository at `https://anonymous.4open.science/r/nn_optimization-82F8/README.md`. Table 1 displays the hyperparameters used in our experimental results. Code is open-sourced at the Github repository.

Table 1: Experiment hyperparameters.

Hyperparameter	**Value**
Batch size	512
Activation function	GeLU
Input data dimension	3072 ($32 \times 32 \times 3$)
Number of epochs	100
α	0.75
Constant learning rates	[0.05, 0.07, 0.1]

Learning rate. To determine a fair range of constant learning rates that balances speed and stability, we use a cyclical learning rate finder as introduced in Smith (2017) and further described in Ahmed (2025). We start with a small learning rate and train the model for a single epoch, repeating this with exponentially increasing learning rates. We plot the loss over the learning rates and pick learning rates in the range where the loss decreases the fastest. Figure 4 displays the result.

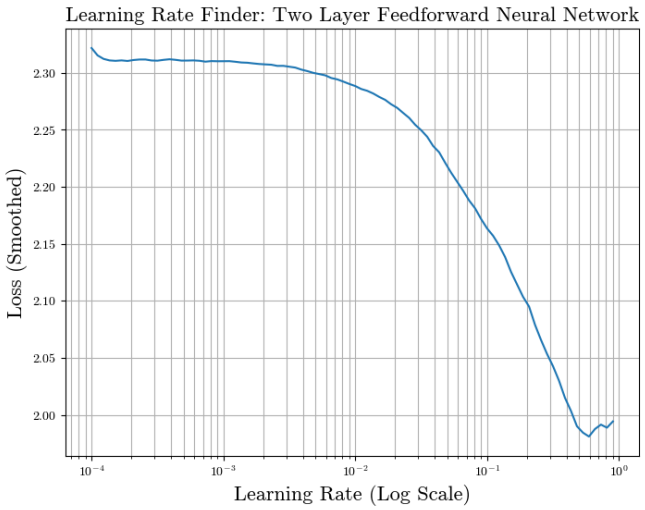

Figure 4: Learning rate selection using a cyclical finder for a two-layer neural network with GELU activations.

Hardware and software. All experiments were run using PyTorch 2.5.0 and CUDA 12.1, on an Intel(R) Xeon(R) Silver 4208 CPU (2.10GHz) with an NVIDIA RTX A6000 GPU.

D.1 FULL COMPARISON BETWEEN CONSTANT AND ADAPTIVE LEARNING RATES

Figure 5 displays the full results of the shared-weight model experiments for a range of L. As discussed in the main body, the adaptive learning rate displays faster and more stable convergence than constant learning rates in the same range of values. Moreover, as the number of layers increase, the iterates undergo a longer initial plateau period, corresponding to a flatter region around initialization.

This further supports the need for a larger learning rate early on to accelerate the iterates out of this region.

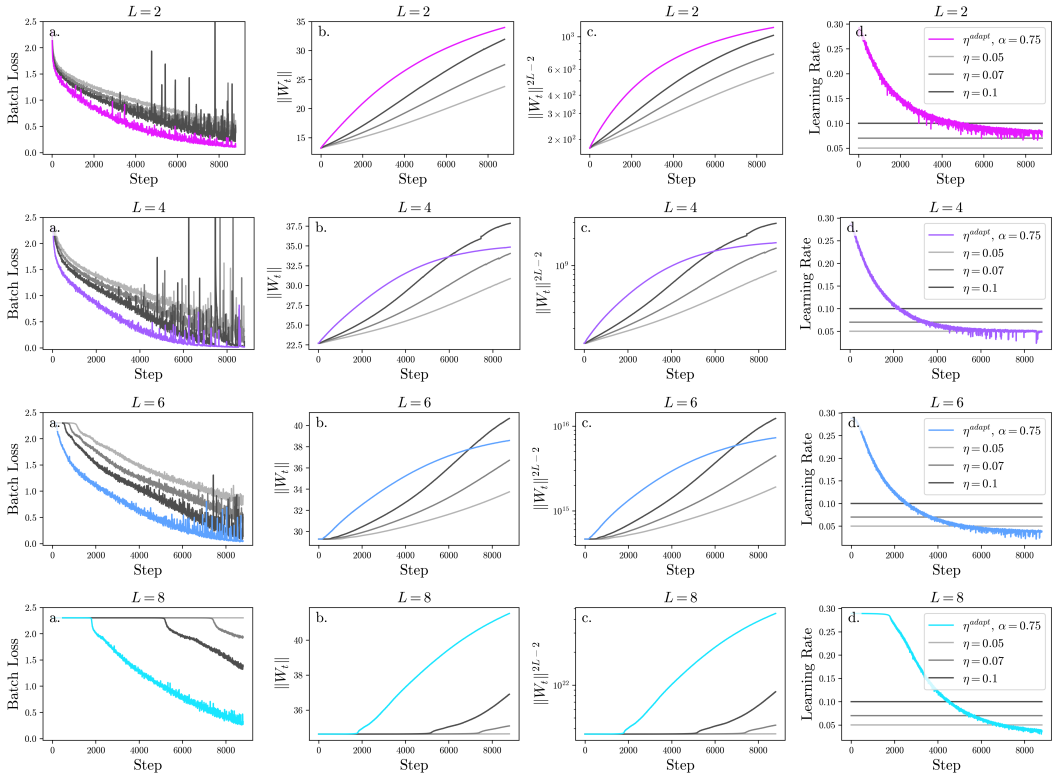

Figure 5: Comparison of constant and adaptive learning rates for standard neural networks of varying depths.