Machine Learning (2022) 111:3359–3392
https://doi.org/10.1007/s10994-022-06192-x

Stabilize deep ResNet with a sharp scaling factor τ

Huishuai Zhang[1] · **Da Yu**[2] · **Mingyang Yi**[3] · **Wei Chen**[4] · **Tie-Yan Liu**[1]

Received: 26 November 2020 / Revised: 10 May 2021 / Accepted: 24 August 2021 /
Published online: 1 August 2022

Abstract
We study the stability and convergence of training deep ResNets with gradient descent. Specifically, we show that the parametric branch in the residual block should be scaled down by a factor $\tau = O(1/\sqrt{L})$ to guarantee stable forward/backward process, where L is the number of residual blocks. Moreover, we establish a converse result that the forward process is unbounded when $\tau > L^{-\frac{1}{2}+c}$, for any positive constant c. The above two results together establish a sharp value of the scaling factor in determining the stability of deep ResNet. Based on the stability result, we further show that gradient descent finds the global minima if the ResNet is properly over-parameterized, which significantly improves over the previous work with a much larger range of τ that admits global convergence. Moreover, we show that the convergence rate is independent of the depth, theoretically justifying the advantage of ResNet over vanilla feedforward network. Empirically, with such a factor τ, one can train deep ResNet without normalization layer. Moreover for ResNets with normalization layer, adding such a factor τ also stabilizes the training and obtains significant performance gain for deep ResNet.

Editor: Paolo Frasconi.

✉ Huishuai Zhang
huzhang@microsoft.com

✉ Wei Chen
chenwei2022@ict.ac.cn

Da Yu
yuda3@mail2.sysu.edu.cn

Mingyang Yi
yimingyang17@mails.ucas.edu.cn

Tie-Yan Liu
tyliu@microsoft.com

[1] Microsoft Research Asia, Beijing, China

[2] Sun Yat-sen University, Guangzhou, China

[3] University of Chinese Academy of Sciences, Beijing, China

[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

1 Introduction

Residual Network (ResNet) has achieved great success in computer vision tasks since the seminal paper (He et al., 2016). The ResNet structure has also been extended to natural language processing and achieved the state-of-the-art performance (Vaswani et al., 2017; Devlin et al., 2018). In this paper, we study the forward/backward stability and convergence of training deep ResNet with gradient descent.

Specifically, we consider the following residual block (He et al., 2016),

$$\text{residual block:} \quad h_l = \phi(h_{l-1} + \tau \mathcal{F}_l(h_{l-1})), \tag{1}$$

where $\phi(\cdot)$ is the point-wise activation function, h_l and h_{l-1} are the output and input of the residual block l, $\mathcal{F}_l(\cdot)$ is the parametric branch, e.g., $\mathcal{F}_l(h_{l-1}) = \boldsymbol{W}_l h_{l-1}$ and $\boldsymbol{W}_l$ is the trainable parameter, and τ is a scaling factor on the parametric branch.

We note that standard initialization schemes, e.g., the Kaiming initialization or the Glorot initialization, are designed to keep the forward and backward variance constant when passing through one layer. However, things become different for ResNet because of the existence of the identity mapping. If $\boldsymbol{W}_l$ adopts the standard initialization, a small τ is necessary for a stable forward process of deep ResNet, because the output magnitude quickly explodes for $\tau = 1$ as L gets large. On the other side, a limit form of *Euler's constant* indicates that $\tau = O(1/L)$ is sufficient for the forward/backward stability, which is assumed in previous work (Allen-Zhu et al., 2018; Du et al., 2019b). We ask

"Are there other values of τ that can guarantee the stability of ResNet with arbitrary depth?"

We answer the above question affirmatively by establishing a non-asymptotic analysis that the stability is guaranteed for deep ResNet with arbitrary depth as long as $\tau = O(1/\sqrt{L})$. Moreover conversely, for any positive constant c, if $\tau = L^{-\frac{1}{2}+c}$, the network output norm grows at least with rate L^c in expectation, which implies the forward/backward process is unbounded as L gets large.

One step further, based on the stability result, we show that if the network is properly over-parameterized, gradient descent finds global minima for training ResNet with $\tau \le \tilde{O}(1/\sqrt{L})$[1]. This is essentially different from previous work that assumes $\tau \le \tilde{O}(1/L)$ (Allen-Zhu et al., 2018; Du et al., 2019a; Frei et al., 2019).

Our contribution is summarized as follows.

- We establish a non-asymptotic analysis showing that $\tau = 1/\sqrt{L}$ is sharp in the order sense to guarantee the stability of deep ResNet.
- For $\tau \le \tilde{O}(1/\sqrt{L})$, we establish the convergence of gradient descent to global minima for training over-parameterized ResNet with a depth-independent rate.

The key step to prove our first claim is a new bound of the spectral norm of the forward process for ResNet with $\tau = O(1/\sqrt{L})$. We find that, although the natural bound $(1 + 1/\sqrt{L})^L$ explodes, the randomness of the trainable parameter in the parametric branch helps to control the output norm growth. Specifically, we bound the mean and the variance about the largest possible change after deep residual mappings when $\tau = O(1/\sqrt{L})$.

[1] We use $\tilde{O}(\cdot)$ to hide logarithmic factors.

We also argue the advantage of adding τ over other stabilization methods, such as *batch normalization* (BN) (Ioffe & Szegedy, 2015) and *Fixup* (Zhang et al., 2018a). First, it has advantage over BN to guarantee stability. BN is architecture-agnostic and the output norm of ResNet with BN still grows unbounded as the depth increases. In practice, it has to to employ a learning rate warm-up stage to train very deep ResNet even with BN (He et al., 2016). In comparison, we prove that ResNet with τ is stable over all depths and hence does not require any learning rate warm-up stage. Second, it is also more stable than the approach of scaling down initialization that is adopted in *Fixup*. Scaling down initial residual weight does not scale down the gradient properly and *Fixup* could explode after gradient descent updates for deep ResNet.

At last, we corroborate the theoretical findings with extensive experiments. First, we demonstrate that with $\tau = 1/\sqrt{L}$, ResNet can be effectively trained without the normalization layers. It is more stable and achieves better performance than *Fixup*. Second, we demonstrate that adding $\tau = 1/\sqrt{L}$ on top of the normalization layer can obtain even better performance.

1.1 Related works

There is a large volume of literature studying ResNet. We can only give a partial list.

To argue the benefit of skip connection, (Veit et al., 2016) interpret ResNet as an ensemble of shallower networks, (Zhang et al., 2018) study the local Hessian of residual blocks, (Hardt & Ma, 2016) show that deep linear residual networks have no spurious local optima, (Orhan & Pitkow, 2018) observe that skip connection eliminates the singularity, and (Balduzzi et al., 2017) find that ResNet is more resistant to the gradient shattering problem than the feedforward network. However, these results mainly rely on empirical observation or strong model assumption.

There are also several papers studying ResNet from the stability perspective (Arpit et al., 2019; Zhang et al., 2018a, b; Yang & Schoenholz, 2017; Haber & Ruthotto, 2017). In comparison, we study the model closest to the original ResNet and provide a rigorous non-asymptotic analysis for the stability when $\tau = O(1/\sqrt{L})$ and a converse result showing the sharpness of τ. We also demonstrate the empirical advantage of learning ResNet with τ.

Our work is also related to recent literature on the theory of learning deep neural network with gradient descent in the over-parameterized regime. Many works (Jacot et al., 2018; Allen-Zhu et al., 2018; Du et al., 2019a; Chizat & Bach, 2018a; Zou et al., 2018; Zou & Gu, 2019; Arora et al., 2019a; Oymak & Soltanolkotabi, 2019; Chen et al., 2019; Ji & Telgarsky, 2019) use Neural Tangent Kernel (NTK) or similar technique to argue the global convergence of gradient descent for training over-parameterized deep neural network. Some (Brutzkus et al., 2017; Li & Liang, 2018; Allen-Zhu et al., 2019a; Arora et al., 2019b; Cao & Gu, 2019; Neyshabur et al., 2019) study the generalization properties of over-parameterized neural network. On the other side, there are papers (Ghorbani et al., 2019; Chizat et al., 2019; Yehudai & Shamir, 2019; Allen-Zhu & Li, 2019) discussing the limitation of the NTK approach in characterizing the behavior of neural network. Additionally, several papers (Chizat & Bach, 2018b; Mei et al., 2018, 2019; Nguyen, 2019; Fang et al., 2019a, a) study the convergence of the weight distribution in the probabilistic space via gradient flow for two or multiple layers network. To the best of our knowledge, we are the first to provide the global convergence of learning ResNet in the regime of $\tau \leq O(1/\sqrt{L})$

2 Preliminaries

There are many residual network models since the seminal paper (He et al., 2016). Here we study a simplified ResNet that shares the same structure as He et al. (2016)[2], which is described as follows,

- Input layer: $h_0 = \phi(\boldsymbol{A}x)$, where $x \in \mathbb{R}^p$ and $\boldsymbol{A} \in \mathbb{R}^{m\times p}$;
- $L-1$ residual blocks: $h_l = \phi(h_{l-1} + \tau \boldsymbol{W}_l h_{l-1})$, where $\boldsymbol{W}_l \in \mathbb{R}^{m\times m}$;
- A fully-connected layer: $h_L = \phi(\boldsymbol{W}_L h_{L-1})$, where $\boldsymbol{W}_L \in \mathbb{R}^{m\times m}$;
- Output layer: $y = \boldsymbol{B}h_L$, where $\boldsymbol{B} \in \mathbb{R}^{d\times m}$;
- Initialization: The entries of $\boldsymbol{A}, \boldsymbol{W}_l$ for $l \in [L-1]$, $\boldsymbol{W}_L$ and $\boldsymbol{B}$ are independently sampled from $\mathcal{N}(0, \frac{2}{m})$, $\mathcal{N}(0, \frac{1}{m})$, $\mathcal{N}(0, \frac{2}{m})$ and $\mathcal{N}(0, \frac{1}{d})$, respectively;

where $\phi(\cdot) := \max\{0, \cdot\}$ is the ReLU activation function. We assume the input dimension is p, the intermediate layers have the same width m and the output has dimension d. For a positive integer L, we use $[L]$ to denote the set $\{1, 2, ..., L\}$. We denote the values before activation by $g_0 = \boldsymbol{A}x, g_l = h_{l-1} + \tau \boldsymbol{W}_l h_{l-1}$ for $l = [L-1]$ and $g_L = \boldsymbol{W}_L h_{L-1}$. We use $h_{i,l}$ and $g_{i,l}$ to denote the value of h_l and g_l, respectively, when the input vector is x_i, and $\boldsymbol{D}_{i,l}$ the diagonal activation matrix where $[\boldsymbol{D}_{i,l}]_{k,k} = \mathbf{1}_{\{(g_{i,l})_k \geq 0\}}$. We use superscript $^{(0)}$ to denote the value at initialization, e.g., $\boldsymbol{W}_l^{(0)}$, $h_{i,l}^{(0)}$, $g_{i,l}^{(0)}$ and $\boldsymbol{D}_{i,l}^{(0)}$. We may omit the subscript $_i$ and the superscript $^{(0)}$ when they are clear from the context for simplifying the notations.

We introduce a notation $\overrightarrow{\boldsymbol{W}} := (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L)$ to represent all the trainable parameters. We note that $\boldsymbol{A}$ and $\boldsymbol{B}$ are fixed after initialization. Throughout the paper, we use $\|\cdot\|$ to denote the l_2 norm of a vector. We further use $\|\cdot\|$ and $\|\cdot\|_F$ to denote the spectral norm and the Frobenius norm of a matrix, respectively. Denote $\|\overrightarrow{\boldsymbol{W}}\| := \max_{l\in[L]} \|\boldsymbol{W}_l\|$ and $\|\boldsymbol{W}_{[L-1]}\| := \max_{l\in[L-1]} \|\boldsymbol{W}_l\|$.

The training data set is $\{(x_i, y_i^*)\}_{i=1}^n$, where x_i is the feature vector and y_i^* is the target signal for $i = 1, ..., n$. We consider the objective function is $F(\overrightarrow{\boldsymbol{W}}) := \sum_{i=1}^n F_i(\overrightarrow{\boldsymbol{W}})$ where $F_i(\overrightarrow{\boldsymbol{W}}) := \ell(\boldsymbol{B}h_{i,L}, y_i^*)$ and $\ell(\cdot)$ is the loss function. The model is trained by running the gradient descent algorithm. Though ReLU is nonsmooth, we abuse the word "gradient" to represent the value computed through back-propagation.

3 Forward and backward stability of ResNet

In this section, we establish the stability of training ResNet. We show that when $\tau = O(1/\sqrt{L})$ the forward and backward pass is bounded at the initialization and after small perturbation. On the converse side, for an arbitrary positive constant c, if $\tau > L^{-0.5+c}$, the output magnitude grows at least polynomial with depth at the initialization. We also argue the advantage of using a factor τ over other stabilization methods, such as BN and *Fixup*. The stability result forms the basis to establish the global convergence in Sect. 4.

[2] In (He et al., 2016), there is a ReLU after the building block $y = x + F(x)$ (please refer to Figure 2 in He et al. (2016)), and hence a whole residual block is $h_l = \phi(h_{l-1} + F(h_{l-1}))$ (if using the notations in our paper).

3.1 Forward process is bounded if $\tau = O(1/\sqrt{L})$

We first give a non-asymptotic bound on the forward process at initialization.

Theorem 1 *Suppose that $\overrightarrow{\boldsymbol{W}}^{(0)}$, $\boldsymbol{A}$ are randomly generated as in the initialization step, and $\boldsymbol{D}_{i,0}^{(0)}, \ldots, \boldsymbol{D}_{i,L}^{(0)}$ are diagonal activation matrices for $i \in [n]$. Suppose that c and ϵ are arbitrary positive constants with $0 < \epsilon < 1$. If τ satisfies $\tau^2 L \le \min\{\frac{1}{2}\log(1+c), \frac{\log^2(1+c)}{16(1+\log(1+2/\epsilon))}\}$, then with probability at least $1 - 3nL^2 \cdot \exp(-m)$ over the initialization randomness, we have for any two integers $a, b \in [L-1]$ with $b > a$ and for all $i \in [n]$,*

$$\left\| \boldsymbol{D}_{i,b}^{(0)} \left(\boldsymbol{I} + \tau \boldsymbol{W}_b^{(0)} \right) \cdots \boldsymbol{D}_{i,a}^{(0)} \left(\boldsymbol{I} + \tau \boldsymbol{W}_a^{(0)} \right) \right\| \le \frac{1+c}{1-\epsilon}. \tag{2}$$

The proof is based on Markov's inequality with recursively conditioning. The full proof is deferred to Appendix B. Here we give an outline.

Proof Outline We omit the subscript i and the superscript (0) for simplicity. Suppose that $\|h_{a-1}\| = 1$. Let $g_l = h_{l-1} + \tau \boldsymbol{W}_l h_{l-1}$ and $h_l = \boldsymbol{D}_l g_l$ for $l = \{a, ..., b\}$. We have

$$\|h_b\|^2 = \frac{\|h_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|h_a\|^2}{\|h_{a-1}\|^2} \|h_{a-1}\|^2 \le \frac{\|g_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|g_a\|^2}{\|h_{a-1}\|^2} \|h_{a-1}\|^2,$$

where the inequality is due to that $\|\boldsymbol{D}_l\| \le 1$. Taking logarithm at both side, we have

$$\log \|h_b\|^2 \le \sum_{l=a}^{b} \log \Delta_l, \quad \text{where } \Delta_l := \frac{\|g_l\|^2}{\|h_{l-1}\|^2}.$$

If let $\tilde{h}_{l-1} := \frac{h_{l-1}}{\|h_{l-1}\|}$, then we obtain that

$$\begin{aligned} \log \Delta_l &= \log\left(1 + 2\tau \left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1} \right\rangle + \tau^2 \|\boldsymbol{W}_l \tilde{h}_{l-1}\|^2\right) \\ &\le 2\tau \left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1} \right\rangle + \tau^2 \|\boldsymbol{W}_l \tilde{h}_{l-1}\|^2, \end{aligned}$$

where the inequality is because $\log(1+x) < x$ for $x > -1$. Let $\xi_l := 2\tau \left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1} \right\rangle$ and $\zeta_l := \tau^2 \|\boldsymbol{W}_l \tilde{h}_{l-1}\|^2$. Then given $\tilde{h}_{l-1}$, we have $\xi_l \sim \mathcal{N}\left(0, \frac{4\tau^2}{m}\right)$, $\zeta_l \sim \frac{\tau^2}{m}\chi_m^2$.

We can argue that $\sum_{l=a}^{b} \xi_l \sim \mathcal{N}\left(0, \frac{4(b-a)\tau^2}{m}\right)$ and $\sum_{l=a}^{b} \zeta_l \sim \frac{(b-a)\tau^2}{m}\chi_m^2$. Hence for arbitrary positive constant c_1, if $\tau^2 L \le c_1/4$ then $\sum_{l=a}^{b} \log \Delta_l \le c_1$ with probability at least $1 - 3\exp(-\frac{mc_1^2}{64\tau^2 L})$. We then convert the condition on c_1 to that on c in the theorem. Taking ϵ -net argument, we can establish the spectral norm bound for all vector h_{a-1}. Let a and b vary from 1 to $L-1$ and taking the union bound gives the claim. The full proof is presented in Appendix B. □

We note that the constant c and ϵ can be chosen arbitrarily small such that $(1+c)/(1-\epsilon)$ is arbitrarily close to 1 given stronger assumption on $\tau^2 L$. Theorem 1 indicates that the norm of every residual block output is upper bounded by $(1+c)/(1-\epsilon)$ if the input vector has norm 1, which demonstrates that the the forward process is stable. This result is a bit surprising since for $\tau = O(1/\sqrt{L})$ a natural bound on the spectral norm $\|(\boldsymbol{I} + \tau \boldsymbol{W}_L^{(0)}) \cdots (\boldsymbol{I} + \tau \boldsymbol{W}_1^{(0)})\| \le (1 + \frac{1}{\sqrt{L}})^L$ explodes. Here the intui-

tion is that the cross-product term concentrates on the mean 0 because of the independent randomness of matrices $\boldsymbol{W}_l^{(0)}$ and the variance can be bounded at the same time.

Moreover, we can also establish a lower bound on the output norm of each residual block as follows.

Theorem 2 *Suppose that c is an arbitrary constant with $0 < c < 1$. If $\tau^2 L \leq \frac{1}{4}\log(1-c)^{-1}$, then with probability at least $1 - 2nL^2 \cdot \exp\left(-\frac{1}{32}m\log(1-c)^{-1}\right)$ over the randomness of $\boldsymbol{A} \in \mathbb{R}^{m\times p}$ and $\overrightarrow{\boldsymbol{W}}^{(0)} \in (\mathbb{R}^{m\times m})^L$ the following holds*

$$\forall i \in [n], l \in [L-1] : \ \left\|h_{i,l}^{(0)}\right\| \geq 1 - c. \tag{3}$$

Proof The proof is similar to that of Theorem 1 but harder. The high level idea is to control the mean and the variance of the mapping of the intermediate residual blocks simultaneously by utilizing the Markov's inequality with the recursive conditioning. The full proof is deferred to Appendix C.1. □

Combining these two theorems, we can conclude that the norm of each residual block output concentrates around 1 with high probability $1 - O(nL^2)\exp(-\Omega(m))$. Moreover these two theorems also holds for $\overrightarrow{\boldsymbol{W}}$ that is within the neighborhood of $\overrightarrow{\boldsymbol{W}}^{(0)}$, which is presented in Appendix C.2.

3.2 Backward process is bounded for $\tau \leq O(1/\sqrt{L})$

For ResNet, the gradient with respect to the parameter is computed through back-propagation. For any input sample i, we denote $\partial h_{i,l} := \frac{\partial F_i(\overrightarrow{\boldsymbol{W}})}{\partial h_{i,l}}$ and $\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}}) := \frac{\partial F_i(\overrightarrow{\boldsymbol{W}})}{\partial \boldsymbol{W}_l} = \left(\tau \boldsymbol{D}_{i,l}\partial h_{i,l}\right) \cdot h_{i,l-1}^T$. Therefore, the gradient upper bound is guaranteed if $h_{i,l}$ and $\partial h_{i,l}$ are bounded for all blocks. We next show that the backward process is bounded for each individual sample at the initialization stage.

Theorem 3 *For every input sample $i \in [n]$ and for any positive constants c and ϵ with $0 < \epsilon < 1$, if τ satisfies $\tau^2 L \leq \min\{\frac{1}{2}\log(1+c), \frac{\log^2(1+c)}{16(1+\log(1+2/\epsilon))}\}$, then with probability at least $1 - 3nL^2 \cdot \exp\left(-\frac{1}{4}mc^2\right)$ over the randomness of $\boldsymbol{A}, \boldsymbol{B}$ and $\overrightarrow{\boldsymbol{W}}^{(0)}$, the following holds $\forall l \in [L-1]$*

$$\|\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}}^{(0)})\|_F \leq \frac{(1+c)^2}{(1-\epsilon)^2}(2\sqrt{2}+c)\tau\|\partial h_{i,L}\|, \ \ \|\nabla_{\boldsymbol{W}_L} F_i(\overrightarrow{\boldsymbol{W}}^{(0)})\|_F \leq \frac{(1+c)}{(1-\epsilon)}\|\partial h_{i,L}\|. \tag{4}$$

The full proof is is deferred to Appendix 6. Here we give an outline.

Proof Outline The argument is based on the back-propagation formula and Theorem 1. We omit the superscript $^{(0)}$ for notation simplicity. For each $i \in [n]$ and $l \in [L-1]$, i.e., the residual layers, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}})\|_F &= \left\|\tau\big(\boldsymbol{D}_{i,l}(\boldsymbol{I}+\tau\boldsymbol{W}_{l+1})^T\cdots\boldsymbol{D}_{i,L-1}\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\partial h_{i,L}\big)h_{i,l-1}^T\right\|_F\\ &\le \tau\|\boldsymbol{D}_{i,l}(\boldsymbol{I}+\tau\boldsymbol{W}_{l+1})^T\cdots\boldsymbol{D}_{i,L-1}\|\cdot\|\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\|\cdot\|\partial h_{i,L}\|\cdot\|h_{i,l-1}\|,\\ &\le \frac{(1+c)^2}{(1-\epsilon)^2}(2\sqrt{2}+c)\tau\|\partial h_{i,L}\|,\end{aligned}$$

where the last inequality is due to Theorem 1 and the spectral norm bound of $\boldsymbol{W}_L$ given in Appendix A. The full proof is deferred to Appendix 6. □

This theorem indicates that the gradient of residual layers could be τ times smaller than the usual feedforward layer. Moreover, for ResNet with $\tau = 1/\sqrt{L}$, the norm of all layer gradient is independent of the depth, which allows us to use a *depth independent learning rate* to train ResNets of all depths. This is essentially different from the feedforward case (Allen-Zhu et al., 2018; Zou & Gu, 2019). We note that the gradient upper bound also holds for $\overrightarrow{\boldsymbol{W}}$ within the neighborhood of $\overrightarrow{\boldsymbol{W}}^{(0)}$ (see details in Appendix C.2 and 6).

3.3 A converse result for $\tau > L^{-\frac{1}{2}+c}$

We have built the stability of the forward/backward process for $\tau = O(1/\sqrt{L})$. We next establish a converse result showing that if τ is slightly larger than $L^{-\frac{1}{2}}$ in the order sense, the network output norm grows uncontrollably as the depth L increases. This justifies the sharpness of the value $\tau = 1/\sqrt{L}$. Without loss of generality, we assume $\|h_0\| = 1$.

Theorem 4 *Suppose that c is an arbitrary positive constant and the ResNet is defined and initialized as in Sect. 2. If $\tau \ge L^{-\frac{1}{2}+c}$, then we have*

$$\mathbb{E}\|h_L\|^2 \ge \frac{1}{2}L^{2c}. \tag{5}$$

Proof The proof is based on a new inequality $(h_l)_k \ge \phi\Big(\sum_{a=1}^{l}\big(\tau\boldsymbol{W}_a h_{a-1}\big)_k\Big)$ for $l \in [L-1]$ and for $k \in [m]$. By the symmetry of Gaussian variables and the recursive conditioning, we can compute the expectation of $\|h_L\|^2$ exactly. The whole proof is relegated to Appendix G. □

This indicates that $\tau = O(1/\sqrt{L})$ is sharp to guarantee the forward stability of deep ResNet. We note that Theorems 1 and 3 hold with high probability when $m > \Omega(\log L)$ and Theorem 2 holds with high probability when $m > \Omega(\log(nL))$. These are very mild conditions on the width m, which are satisfied by practical networks.

3.4 Comparison with other approaches for stability

Up to now, we have provided a sharp value of τ in terms of determining the stability of deep ResNet. In practice, two other approaches are used in residual networks to provide the stability: adding normalization layers, e.g., batch normalization (BN) (Ioffe & Szegedy, 2015), and scaling down the initial residual weights, e.g., *Fixup* (Zhang et al., 2018a). Next, we discuss BN and *Fixup* from the stability perspective, respectively, and make comparison with adding $\tau = 1/\sqrt{L}$.

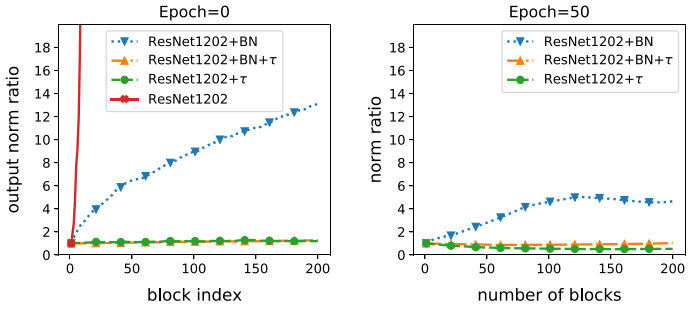

Fig. 1 The l_2 norm of residual block output of the first stage of ResNet1202 at epoch 0 and epoch 50. The X axis is the block index and the Y axis is the output norm ratio compared to the first block

Batch normalization is placed right after each convolutional layer in (He et al., 2016). Here for the ResNet model defined in Sect. 2, we put BN after each parametric branch and the residual block becomes $h_l = \phi(h_{l-1} + \tilde{z}_l)$, where $(\tilde{z}_{i,l})_k := \mathrm{BN}\big((z_{i,l})_k\big) = \frac{(z_{i,l})_k - \mathbb{E}[(z_{\cdot,l})_k]}{\sqrt{\mathrm{Var}[(z_{\cdot,l})_k]}}$ and $(z_{i,l})_k := \big(\boldsymbol{W}_l h_{i,l-1}\big)_k$ for $k = [m]$ and $l = [L-1]$, and the expectation and the variance are taken over samples in a mini-batch. Then we have $\mathbb{E}(\tilde{z}_{\cdot,l})_k = 0$ and $\mathrm{Var}[(\tilde{z}_{\cdot,l})_k] = 1$. We use the following proposition to estimate the norm of each residual block output for the ResNet with BN.

Proposition 1 *Assume that $(\tilde{z}_l)_k$ are independent random variable over l, k with $\mathbb{E}(\tilde{z}_l)_k = 0$ and $\mathrm{Var}[(\tilde{z}_l)_k] = 1$. The output norm of the residual block l satisfies $\mathbb{E}\|h_l\|^2 \geq \frac{1}{2}ml$, for $l \in [L-1]$.*

Proof The proof is adapted from the proof of Theorem 4, and is presented in Appendix G. □

This indicates that the block output norm of ResNet with BN grows roughly at the rate $\sqrt{l}$ at the initialization stage, where l is the block index and the larger l the closer to the output. To verify this, we plot how the output norm of each residual block grows for ResNet1202 (with/without BN)[3] in Fig. 1. We see that at epoch 0 (initialization stage), the output norm grows almost with the rate $\sqrt{l}$ as predicted in Proposition 1. After training, the estimation in Proposition 1 is not as accurate as the initialization because the independence assumption does not hold after training. Besides the output norm growth, in practice, (He et al., 2016) have to use warm-up learning rates to train very deep ResNets, e.g., ResNet1202+BN. In contrast, it is proved that the approach of adding $\tau = 1/\sqrt{L}$ is stable over all depths and hence does not require any learning rate warm-up stage.

Recently, Zhang et al. (2018a) propose *Fixup* to train residual networks without the normalization layer. Essentially for each residual block, *Fixup* sets the weight matrix near the output to be 0 at the initialization stage, and then scales down the all other weight matrices

[3] Throughout the paper, the naming rule of ResNet is as follows."ResNet" is referred to the model defined in Sect. 2, "ResNet#" is referred to the models in He et al. (2016) with removing all the BN layers, e.g., ResNet1202, "ResNet#+BN" corresponds to the original model in He et al. (2016), "+Fixup" corresponds to initializing the model with Fixup, and "$+\tau$" is referred to adding τ on the output of the parametric branch in each residual block.

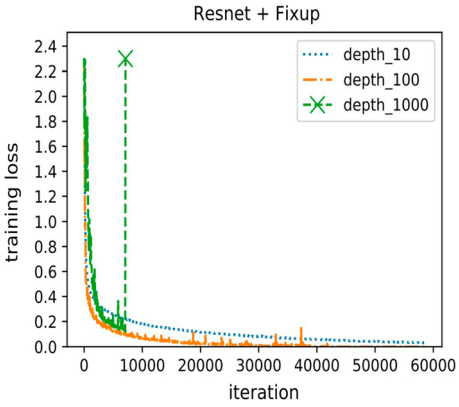

Fig. 2 Training curves of ResNets with Fixup: MNIST classification, width $m = 128$ and learning rate $\eta = 0.01$

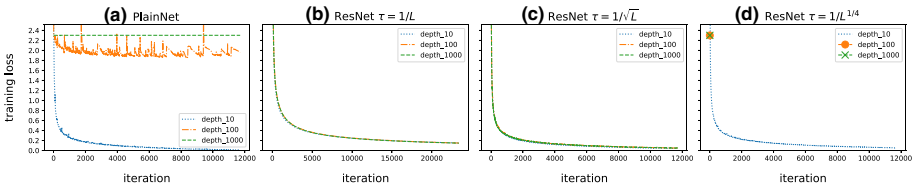

Fig. 3 Training curves for PlainNet, ResNet with $\tau = \frac{1}{L}$, $\tau = \frac{1}{\sqrt{L}}$ and $\tau = \frac{1}{L^{1/4}}$ (from left to right). We use markers to denote the training encounters numerical overflow

by a factor that is determined by the network structure. However, in practice *Fixup* does not always converge for training very deep residual networks as shown in Sect. 5.2. Moreover, for the ResNet model defined in Sect. 2, *Fixup* could be unstable after gradient updates. The residual block is given by $h_l = \phi(h_{l-1} + \boldsymbol{W}_l h_{l-1})$, and following *Fixup*, $\boldsymbol{W}_l^{(0)}$ is initialized to be 0 for $l \in [L-1]$. At the initial stage for input sample i, $h_{i,l} = h_{i,0}$ and hence $\nabla_{\boldsymbol{W}_l} F_i = \partial h_{i,L-1} \cdot h_{i,0}^T$, the same for all $l \in [L-1]$. Then after one gradient update the residual blocks mapping $\prod_{l=1}^{L-1} \boldsymbol{D}_{i,l}(\boldsymbol{I} + \eta \cdot \nabla_{\boldsymbol{W}_l} F_i)$ could behave like $(\boldsymbol{D}(\boldsymbol{I} + \eta \cdot \partial h_{i,L-1} \cdot h_{i,0}^T))^{L-1}$ when $\boldsymbol{D}_{i,l} = \boldsymbol{D}$ for all l, which grows exponentially. Empirically, such explosion is observed for deep ResNet with *Fixup* (see Fig. 2). In contrast, the ResNet with τ is stable for varying depths (see Fig. 3), as guaranteed by our theory.

4 Global convergence for over-parameterized ResNet

In this section, we establish that gradient descent converges to global minima for learning an over-parameterized ResNet with $\tau \leq \tilde{O}(1/\sqrt{L})$. Compared to the recent work (Allen-Zhu et al., 2018), our result significantly enlarges the region of τ that admits the global convergence of gradient descent. Moreover, our result also theoretically justifies the advantage of ResNet over vanilla feedforward network in terms of facilitating the convergence of gradient descent. Before stating the theorem, we introduce common assumptions on the training data and the loss function (Allen-Zhu et al., 2018; Zou & Gu, 2019; Oymak & Soltanolkotabi, 2019).

Assumption 1 (training data) For any x_i, it holds that $\|x_i\| = 1$ and $(x_i)_p = 1/\sqrt{2}$. There exists $\delta > 0$, such that $\forall i, j \in [n], i \neq j, \|x_i - x_j\| \geq \delta$.

The loss function $\ell(\cdot, \cdot)$ is quadratic and the individual objective is $F_i(\vec{\boldsymbol{W}}) := \frac{1}{2}\|\boldsymbol{B}h_{i,L} - y_i^*\|^2$. We note that the assumption $(x_i)_p = 1/\sqrt{2}$ means that the last coordinate of every x_i is $1/\sqrt{2}$. This gives a random bias term after the first layer $\boldsymbol{A}(\cdot)$, which makes the proof of Lemma 6 for the gradient lower bound easier. This assumption is because of the proof convenience rather than something that should be satisfied in practice.

Theorem 5 *Suppose that the ResNet is defined and initialized as in Sect. 2 with* $\tau \leq O(1/(\sqrt{L}\log m))$ *and the training data satisfy Assumption 1. If the network width* $m \geq \Omega(n^8 L^7 \delta^{-4} d \log^2 m)$, *then with probability at least* $1 - \exp(-\Omega(\log^2 m))$, *gradient descent with learning rate* $\eta = \Theta(\frac{d}{nm})$ *finds a point* $F(\vec{\boldsymbol{W}}) \leq \varepsilon$ *in* $T = \Omega(n^2\delta^{-1}\log\frac{n\log^2 m}{\varepsilon})$ *iterations.*

Proof The full proof is deferred to Appendix F. □

This theorem establishes the linear convergence of gradient descent for learning ResNet for the range $\tau \leq O(1/(\sqrt{L}\log m))$. Combined with the unstable case of $\tau > 1/\sqrt{L}$ in Sect. 3.3, we give a nearly full characterization of the convergence in terms of the range of τ. Moreover, our result indicates that the learning rate and the total number of iterations are depth-independent. We note that a recent paper Frei et al. (2019) also achieves a depth-independent rate but only for the case $\tau \leq O(1/(L\log m))$, whose proof critically relies on the choice of $\tau = 1/L$. The overparameterization dependence and the number of iterations are not directly comparable as we are studying the regression problem while Frei et al. (2019) is for the classification problem with different data assumption. Other previous results (Allen-Zhu et al., 2018; Du et al., 2019a) characterize the convergence guarantee only for the case $\tau \leq O(1/(L\log m))$, and their total number of iterations scales with the order L^2. Our depth-independent results are achieved by a tighter smoothness and gradient upper bound.

In the analysis with the feedforward case (Allen-Zhu et al., 2018; Zou & Gu, 2019, the learning rate has to scale with $1/L^2$ and the total number of iterations scales with L^2 for the convergence of learning feedforward network. Therefore, our result theoretically justifies the advantage of ResNet over vanilla feedforward network in terms of facilitating the convergence of gradient descent.

Finally, we add a remark on the width requirement in Theorem 5. The width grows polynomially with the number of training examples. Such dependence is because we need to more neurons to distinguish each data point sufficiently with more examples, which is common for the regression task (Allen-Zhu et al., 2019b; Zou & Gu, 2019). This dependence could be avoid by assuming the training data follows specific distributions for the classification task (Cao & Gu, 2020). However this is orthogonal to our main claim that ResNet converges with a *depth-dependent* rate.

5 Empirical study

In this section, we present experiments to verify our theory and show the practical value of ResNet with τ. We first compare the performance of ResNet with different τ's and demonstrate that $\tau = \frac{1}{\sqrt{L}}$ is a sharp value in determining the trainability of deep ResNet. We then

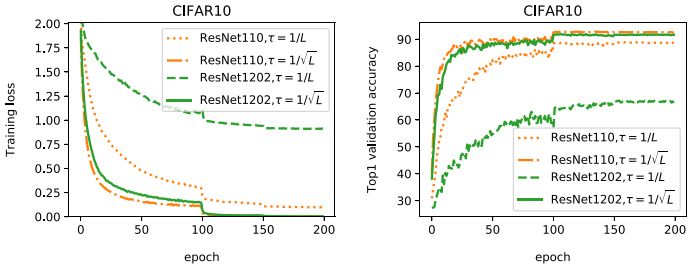

Fig. 4 Training/validation curves of ResNet110/1202 with $\tau = 1/\sqrt{L}$ and $\tau = 1/L$ for CIFAR10 classification task. We use the models in He et al. (2016) and remove all BN layers

compare the performance of adding the factor τ and using *Fixup* initialization when training the popular residual networks without normalization layers. We finally show that with normalization layer, adding τ also significantly improve the performance for both CIFAR and ImageNet tasks. Source code available online https://github.com/dayu11/tau-ResNet.

5.1 Theoretical verification

We train feedforward fully-connected neural networks (PlainNet), ResNets with different values of τ, and compare their convergence behaviors. Specifically, for ResNets, we adopts the exactly the same residual architecture as described in Eq. (1) and Sect. 2. The PlainNet adopts the same architecture as the ResNets without the skip connection. The models are generated with width $m = 128$ and depth $L \in \{10, 100, 1000\}$. For ResNets with τ, we choose $\tau = \frac{1}{L}, \frac{1}{\sqrt{L}}, \frac{1}{L^{1/4}}$ to show the sharpness of the value $\frac{1}{\sqrt{L}}$. We conduct classification on the MNIST dataset (LeCun et al., 1998). We train the model with SGD[4] and the size of minibatch is 256. The learning rate is set to $\eta = 0.01$ for all networks without tuning.

We plot the training curves in Fig. 3. For ResNets with τ, we see that both $\tau = \frac{1}{L}$ and $\tau = \frac{1}{\sqrt{L}}$ are able to train very deep ResNets successfully and $\tau = \frac{1}{\sqrt{L}}$ achieves lower training loss than $\tau = \frac{1}{L}$. For $\tau = \frac{1}{L^{1/4}}$, the training loss explodes for models with depth 100 and 1000. This indicates that the bound $\tau = \frac{1}{\sqrt{L}}$ is sharp for learning deep ResNets. Moreover, the convergence of ResNets with $\tau = \frac{1}{\sqrt{L}}$ does not depend on the depth while training feedforward network becomes harder as the depth increases, corroborating our theory nicely.

To clearly see the benefit of $\tau = \frac{1}{\sqrt{L}}$ over $\tau = \frac{1}{L}$, we conduct the classification task on the CIFAR10 dataset (Krizhevsky & Hinton, 2009) with the residual networks from He et al. (2016). A bit different from the model described in Sect. 2, here one residual block is composed of two stacked convolution layers. We argue that our theoretical analysis still applies if treating the number of channels in convolution layer as width in Sect. 2. We plot the training/validation curves in Fig. 4. We can see that with $\tau = \frac{1}{\sqrt{L}}$, both ResNet110 and ResNet1202 can be trained to good accuracy without BN. In contrast, with $\tau = \frac{1}{L}$, the performance of ResNet110 and ResNet1202 drops a lot.

In the sequel, we use "adding τ^*" or "$+\tau^*$" to denote residual network with $\tau = \frac{1}{\sqrt{L}}$.

[4] GD exhibits the same phenomenon. We use SGD due to the expensive per-iteration cost of GD.

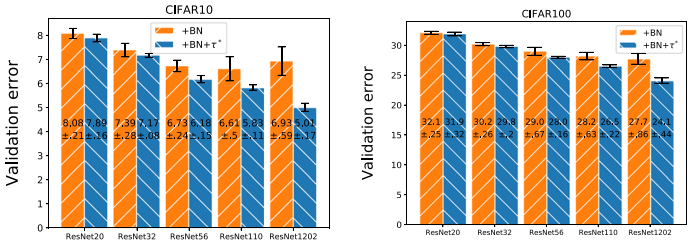

Fig. 5 Validation error bar charts for CIFAR classification tasks. Numbers are average of 5 runs with standard deviations. The deeper network, the larger benefit of τ^*

Table 1 Validation errors of ResNets+Fixup and ResNets+τ^* on CIFAR10. Numbers are average of 5 runs with standard deviations

Model	+ Fixup	+ τ^*
ResNet20	8.72(±0.26)	**8.39** (±0.11)
ResNet32	7.99(±0.24)	**7.68** (±0.10)
ResNet110	7.24(±0.12)	**6.52** (±0.20)
ResNet1202	7.83(N/A)	**6.08** (±0.21)

The relatively better error rates are in bold

5.2 Comparison of adding τ^* and using *Fixup*

In this section we compare our approach of adding τ^* and the approach of using *Fixup* for training residual networks without BN. We conduct the classification task on the CIFAR10 dataset. We use the residual models in (He et al., 2016) with removing all the normalization layers. For the approach of *Fixup*, we use the code from their github website with the same hyperparameter setting. We note that *Fixup* has a learnable scalar with initial value 1 on the output of the parametric branch in each residual block, which is equivalent to set $\tau = 1$. For our approach, we use the same model as *Fixup* with setting $\tau = \frac{1}{\sqrt{L}}$ and using the Kaiming initialization instead of *Fixup* initialization.

The results are presented in Table 1. We can see that our approach achieves much better performance than the *Fixup* approach over all depths. Moreover, the *Fixup* approach fails to converge 2 out of 5 runs for training ResNet1202 and hence the standard deviation is not presented in Table 1.

5.3 Add τ^* on top of normalization

In this section, we empirically show that adding τ^* in the residual block with batch normalization can also help to achieve better performance. We conduct experiments on standard classification datasets: CIFAR10/100 and ImageNet. The baseline models are the residual networks in He et al. (2016). We note that the residual block here is with batch normalization, which is discussed in Sect. 3.4 but not precisely covered by the theoretical model (Sect. 2). For our approach, the only modification is adding a fixed $\tau = \frac{1}{\sqrt{L}}$ at the output of each residual block (right before the residual addition). We also tried to use learnable τ but did not observe gain, which may be due to that the BN layers have learnable

Table 2 Top1 validation error on ImageNet. The models are adapted from He et al. (2016)

Model	Method	Error
ResNet50	+BN	23.6
	+BN+τ^*	**22.7**
ResNet101	+BN	22.0
	+BN+τ^*	**21.4**
ResNet152	+BN	21.7
	+BN+τ^*	**20.9**

The relatively better error rates are in bold

scaling factors. The validation errors on CIFAR10/100 are illustrated in Fig. 5, where all numbers are averaged over five runs. The performance of adding τ^* is much better than the baseline models and especially the benefit of adding τ^* becomes larger when the network is deeper. We note that one needs warm-up learning rate to successfully train ResNet1202+BN, while with τ^* we use the same learning rate schedule for all depths.

As the models for ImageNet classification has different numbers of residual blocks in each stage, we choose $\tau^* = \frac{1}{\sqrt{L}}$ where L is the average number of blocks over all stages. We take average instead of sum because there exists a BN layer on the output of each stage. All models are trained for 200 epochs with learning rate divided by 10 every 60 epochs. The other hyperparameters are the same as in He et al. (2016). Table 2 shows the top 1 validation error results on ImageNet. We can see that just by adding τ^* on top of BN we can achieve significant performance gain.

6 Conclusion

In this paper, we provide a non-asymptotic analysis on the forward/backward stability for ResNet, which unveils that $\tau = 1/\sqrt{L}$ is a sharp value in terms of characterizing the stability. We also bridge theoretical understanding and practical guide of ResNet structure. We empirically verify the efficacy of adding τ for ResNet with/without batch normalization. As the residual block is also widely used in the *Transformer* model (Vaswani et al., 2017), it is interesting to study the effect of τ and layer normalization there.

A Useful Lemmas

First we list several useful bounds on Gaussian distribution.

Lemma 1 *Suppose* $X \sim \mathcal{N}(0, \sigma^2)$, *then*

$$\begin{aligned} &\mathbb{P}\{|X| \le x\} \ge 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right), \\ &\mathbb{P}\{|X| \le x\} \le \sqrt{\frac{2}{\pi}}\frac{x}{\sigma}. \end{aligned} \tag{6}$$

Another bound is on the spectral norm of random matrix ((Vershynin, 2012), Corollary 5.35).

Lemma 2 *Let $\boldsymbol{A} \in \mathbb{R}^{N\times n}$, and entries of $\boldsymbol{A}$ are independent standard Gaussian random variables. Then for every $t \geq 0$, with probability at least $1 - \exp(-t^2/2)$ one has*

$$s_{\max}(\boldsymbol{A}) \leq \sqrt{N} + \sqrt{n} + t, \tag{7}$$

where $s_{\max}(\boldsymbol{A})$ are the largest singular value of $\boldsymbol{A}$.

B Spectral norm bound at initialization

Next we present a spectral norm bound related to the forward process of ResNet with τ.

Proof Without introducing ambiguity, we drop the superscript $^{(0)}$ for notation simplicity. We first build the claim for one fixed sample $i \in [n]$ and drop the subscript i, for convenience. Let $g_l = h_{l-1} + \tau \boldsymbol{W}_l h_{l-1}$ and $h_l = \boldsymbol{D}_l g_l$ for $l = \{a, ..., b\}$. We will show for a vector h_{a-1} with $\|h_{a-1}\| = 1$, we have $\|h_b\| \leq 1 + c$ with high probability, where

$$h_b = \boldsymbol{D}_b(\boldsymbol{I} + \tau \boldsymbol{W}_b)\boldsymbol{D}_{b-1} \cdots \boldsymbol{D}_a(\boldsymbol{I} + \tau \boldsymbol{W}_a)h_{a-1}. \tag{8}$$

Then we have $\|g_l\| \geq \|h_l\|$ due to the assumption $\|\boldsymbol{D}_l\| \leq 1$. Hence we have

$$\|h_b\|^2 = \frac{\|h_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|h_a\|^2}{\|h_{a-1}\|^2}\|h_{a-1}\|^2 \leq \frac{\|g_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|g_a\|^2}{\|h_{a-1}\|^2}\|h_{a-1}\|^2.$$

Taking logarithm at both side, we have

$$\log \|h_b\|^2 \leq \sum_{l=a}^{b} \log \Delta_l, \qquad \text{where } \Delta_l := \frac{\|g_l\|^2}{\|h_{l-1}\|^2}. \tag{9}$$

If letting $\tilde{h}_{l-1} := \frac{h_{l-1}}{\|h_{l-1}\|}$, then we obtain that

$$\begin{aligned} \log \Delta_l &= \log\left(1 + 2\tau\left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1}\right\rangle + \tau^2 \|\boldsymbol{W}_l \tilde{h}_{l-1}\|^2\right) \\ &\leq 2\tau\left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1}\right\rangle + \tau^2 \|\boldsymbol{W}_l \tilde{h}_{l-1}\|^2, \end{aligned}$$

where the inequality is due to the fact $\log(1 + x) \leq x$ for all $x > -1$. Let $\xi_l := 2\tau\left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l \tilde{h}_{l-1}\right\rangle$ and $\zeta_l := \tau^2 \|\boldsymbol{W}_l^{(0)} \tilde{h}_{l-1}\|^2$, then given h_{l-1} we have $\xi_l \sim \mathcal{N}\left(0, \frac{4\tau^2}{m}\right)$, $\zeta_l \sim \frac{\tau^2}{m}\chi_m^2$ because of the random initialization of $\boldsymbol{W}_l$. We see that

$$\mathbb{P}\left(\sum_{l=a}^{b} \log \Delta_l \geq c_1\right) \leq \mathbb{P}\left(\sum_{l=a}^{b} \xi_l \geq \frac{c_1}{2}\right) + \mathbb{P}\left(\sum_{l=a}^{b} \zeta_l \geq \frac{c_1}{2}\right). \tag{10}$$

Next we bound the two terms on the right hand side one by one. For the first term we have

$$\mathbb{P}\left(\sum_{l=a}^{b}\xi_l \geq \frac{c_1}{2}\right) = \mathbb{P}\left(\exp\left(\lambda\sum_{l=a}^{b}\xi_l\right) \geq \exp\left(\frac{\lambda c_1}{2}\right)\right) \leq \mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b}\xi_l - \frac{\lambda c_1}{2}\right)\right], \tag{11}$$

where λ is any positive number and the last inequality uses the Markov's inequality. Moreover,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b}\xi_l\right)\right] &= \mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b-1}\xi_l\right)\mathbb{E}\left[\exp\left(\lambda\xi_b\right)\right]\Big|\mathcal{F}_{b-1}\right] \\ &= \exp\left(\frac{4\tau^2\lambda^2}{m}\right)\mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b-1}\xi_l\right)\right] \\ &= \cdots = \exp\left(\frac{4\tau^2\lambda^2(b-a+1)}{m}\right).\end{aligned} \tag{12}$$

Hence we obtain

$$\mathbb{P}\left(\sum_{l=a}^{b}\xi_l \geq \frac{c_1}{2}\right) \leq \exp\left(\frac{4m^2c_1^2\tau^2(b-a+1)}{256m\tau^4L^2} - \frac{mc_1^2}{32\tau^2L}\right) = \exp\left(-\frac{mc_1^2}{64\tau^2L}\right), \tag{13}$$

by choosing $\lambda = \frac{mc_1}{16\tau^2L}$ and using $b-a+1 \leq L$. Due to the symmetry of $\sum_{l=a}^{b}\xi_l$, the conclusion can be generalized to the quantity $|\sum_{l=a}^{b}\xi_l|$ that $\mathbb{P}\left(\left|\sum_{l=a}^{b}\xi_l\right| \geq \frac{c_1}{2}\right) \leq 2\exp\left(-\frac{mc_1^2}{64\tau^2L}\right)$.

Then, for the second term, we follow the above procedure but for a χ_m^2 variable. We note that the generate moment function of χ_m^2 is $(1-2t)^{-m/2}$ for $t < 1/2$. We will use an inequality that $(1-\frac{x}{m})^{-m} \leq e^x$ for $x \geq 0$. By using the Markov's inequality, we first have for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{l=a}^{b}\zeta_l \geq \frac{c_1}{2}\right) = \mathbb{P}\left(\exp\left(\lambda\sum_{l=a}^{b}\zeta_l\right) \geq \exp\left(\frac{\lambda c_1}{2}\right)\right) \leq \mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b}\zeta_l - \frac{\lambda c_1}{2}\right)\right]. \tag{14}$$

Then we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b}\zeta_l\right)\right] &= \mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b-1}\zeta_l\right)\mathbb{E}\left[\exp\left(\lambda\zeta_b\right)\right]\Big|\mathcal{F}_{b-1}\right] \\ &= \left(1-\frac{\lambda\tau^2}{m/2}\right)^{-m/2}\mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b-1}\zeta_l\right)\right] \\ &\leq \exp(\lambda\tau^2)\mathbb{E}\left[\exp\left(\lambda\sum_{l=a}^{b-1}\zeta_l\right)\right] \\ &\leq \cdots \leq \exp\left(\lambda\tau^2(b-a+1)\right).\end{aligned} \tag{15}$$

Hence we obtain

$$\mathbb{P}\left(\sum_{l=a}^{b}\zeta_l \geq \frac{c_1}{2}\right) \leq \exp\left(\lambda\tau^2(b-a+1) - \frac{\lambda c_1}{2}\right) \leq \exp\left(-\frac{mc_1^2}{2\tau^2 L}\left(1-\frac{2\tau^2 L}{c_1}\right)\right), \tag{16}$$

by choosing $\lambda = \frac{mc_1}{\tau^2 L}$ and using $b-a+1 \leq L$. If further setting τ such that $\tau^2 L \leq \frac{c_1}{4}$, we have

$$\mathbb{P}\left(\sum_{l=a}^{b}\zeta_l \geq \frac{c_1}{2}\right) \leq \exp\left(-\frac{mc_1^2}{4\tau^2 L}\right). \tag{17}$$

Combining (13) and (17), we obtain $\mathbb{P}\left(\sum_{l=a}^{b}\log\Delta_l \geq c_1\right) \leq 3\exp\left(-\frac{mc_1^2}{64\tau^2 L}\right)$ under the condition $\tau^2 L \leq \frac{c_1}{4}$. Hence we have $\mathbb{P}\big(\|h_b\| \geq 1+c\big) \leq \mathbb{P}\left(\sum_{l=a}^{b}\log\Delta_l \geq 2\log(1+c)\right) \leq 3\exp\left(-\frac{m\log^2(1+c)}{16\tau^2 L}\right)$ under the condition that $\tau^2 L \leq \frac{1}{2}\log(1+c)$. We next use ϵ-net argument to prove the claim for all m-dimensional vectors of h_{a-1}. Let $\mathcal{N}_\epsilon$ be an ε-net over the unit ball in $\mathbb{R}^m$ with $\epsilon < 1$, then we have the cardinality $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^m$. Taking the union bound over all vectors h_{a-1} in the net $\mathcal{N}_\epsilon$, we obtain

$$\begin{aligned}\mathbb{P}\left\{\max_{h_{a-1}\in\mathcal{N}_\epsilon}\|h_b\| > 1+c\right\} &\leq (1+2/\epsilon)^m \cdot 3\exp\left(-\frac{m\log^2(1+c)}{16\tau^2 L}\right)\\ &= 3\exp\left(-m\left(\frac{\log^2(1+c)}{16\tau^2 L} - \log(1+2/\epsilon)\right)\right) \leq 3\exp(-m),\end{aligned}$$

where the last equality is obtained by choosing τ appropriately to make $\frac{\log^2(1+c)}{16\tau^2 L} - \log(1+2/\epsilon) > 1$. Then we have the spectral norm bound

$$\left\|\boldsymbol{D}_b\left(\boldsymbol{I}+\tau\boldsymbol{W}_b^{(0)}\right)\boldsymbol{D}_{b-1}\cdots\boldsymbol{D}_a\left(\boldsymbol{I}+\tau\boldsymbol{W}_a^{(0)}\right)\right\| \leq (1-\epsilon)^{-1}\max_{h_{a-1}\in\mathcal{N}_\epsilon}\|h_b\|.$$

This is because of the following argument. For a matrix $\boldsymbol{M}$, v_i is a vector in the net which is closest to a unit vector v, then $\|\boldsymbol{M}v\| \leq \|\boldsymbol{M}v_i\| + \|\boldsymbol{M}(v-v_i)\| \leq \|\boldsymbol{M}v_i\| + \epsilon\|\boldsymbol{M}\|$, and hence taking the supremum over v, one obtains $(1-\epsilon)\|\boldsymbol{M}\| \leq \max_i\|\boldsymbol{M}v_i\|$.

Finally taking a union bound over a and b with $1 \leq a \leq b < L$ and a union bound over all samples $i \in [n]$, we have the claimed result. □

C Bounded forward/backward process

C.1 Proof at initialization

Proof We ignore the subscript $^{(0)}$ for simplicity. First we have

$$\|h_{i,l}\| = \|h_{i,0}\|\frac{\|h_{i,1}\|}{\|h_{i,0}\|}\cdots\frac{\|h_{i,l}\|}{\|h_{i,l-1}\|}. \tag{18}$$

Then we see

$$\log \|h_{i,l}\|^2 = \log \|h_{i,0}\|^2 + \sum_{a=1}^{l} \log \frac{\|h_{i,a}\|^2}{\|h_{i,a-1}\|^2} = \log \|h_{i,0}\|^2 + \sum_{a=1}^{l} \log \left(1 + \frac{\|h_{i,a}\|^2 - \|h_{i,a-1}\|^2}{\|h_{i,a-1}\|^2}\right). \tag{19}$$

We introduce notation $\Delta_a := \frac{\|h_{i,a}\|^2 - \|h_{i,a-1}\|^2}{\|h_{i,a-1}\|^2}$. We next give a lower bound on Δ_a. Let S be the set $\{k : k \in [m] \text{ and } (h_{i,a-1})_k + \tau(\boldsymbol{W}_a h_{i,a-1})_k > 0\}$. We have that

$$\begin{aligned}
\Delta_a &= \frac{1}{\|h_{i,a-1}\|^2} \sum_{k \in S} \left[(h_{i,a-1})_k^2 + 2\tau(h_{i,a-1})_k(\boldsymbol{W}_a h_{i,a-1})_k + (\tau \boldsymbol{W}_a h_{i,a-1})_k^2\right] - \frac{1}{\|h_{i,a-1}\|^2} \sum_{k=1}^{m} (h_{i,a-1})_k^2 \\
&= -\frac{1}{\|h_{i,a-1}\|^2} \sum_{k \notin S} (h_{i,a-1})_k^2 + \frac{1}{\|h_{i,a-1}\|^2} \sum_{k \in S} \tau^2 (\boldsymbol{W}_a h_{i,a-1})_k^2 + \frac{2}{\|h_{i,a-1}\|^2} \sum_{k \in S} \tau (h_{i,a-1})_k (\boldsymbol{W}_a h_{i,a-1})_k \\
&\geq -\frac{1}{\|h_{i,a-1}\|^2} \sum_{k=1}^{m} (\tau \boldsymbol{W}_a h_{i,a-1})^2 + \frac{2}{\|h_{i,a-1}\|^2} \tau \sum_{k=1}^{m} (h_{i,a-1})_k (\boldsymbol{W}_a h_{i,a-1})_k \\
&= -\frac{\|\tau \boldsymbol{W}_a h_{i,a-1}\|^2}{\|h_{i,a-1}\|^2} + \frac{2\tau \langle h_{i,a-1}, \boldsymbol{W}_a h_{i,a-1} \rangle}{\|h_{i,a-1}\|^2},
\end{aligned} \tag{20}$$

where the inequality is due to the fact that for $k \notin S$, $|(h_{i,a-1})_k| < |(\tau \boldsymbol{W}_a h_{i,a-1})_k|$ and $(h_{i,a-1})_k (\boldsymbol{W}_a h_{i,a-1})_k \leq 0$. Let $\xi_a := \frac{2\tau \langle h_{i,a-1}, \boldsymbol{W}_a h_{i,a-1} \rangle}{\|h_{i,a-1}\|^2}$ and $\zeta_a := \frac{\|\tau \boldsymbol{W}_a h_{i,a-1}\|^2}{\|h_{i,a-1}\|^2}$, then $\Delta_a \geq \xi_a - \zeta_a$. We note that given $h_{i,a-1}$, $\xi_a \sim \mathcal{N}\left(0, \frac{4\tau^2}{m}\right)$ and $\zeta_a \sim \frac{\tau^2}{m} \chi_m^2$. We use a tail bound for a χ_m^2 variable X (see Lemma 1 in Laurent and Massart (2000))

$$\mathbb{P}(|X - m| \geq u) \leq e^{-\frac{u^2}{4m}}. \tag{21}$$

By applying the tail bound on Gaussian and Chi-square variables, for a constant c_0 such that $4\tau^2 \leq c_0$ we have

$$\begin{aligned}
\mathbb{P}\left(\Delta_a < -c_0\right) &= \mathbb{P}\left(\Delta_a < -c_0 \text{ and } \xi_a < -\frac{c_0}{2}\right) + \mathbb{P}\left(\Delta_a < -c_0 \text{ and } \xi_a \geq -\frac{c_0}{2}\right) \\
&\leq \mathbb{P}\left(\xi_a < -\frac{c_0}{2}\right) + \mathbb{P}\left(\zeta_a > \frac{c_0}{2}\right) \\
&= \frac{1}{2} \exp\left(-\frac{mc_0^2}{32\tau^2}\right) + \exp\left(-\frac{mc_0^2}{16\tau^4}\right) \\
&< \exp\left(-\frac{mc_0^2}{32\tau^2}\right).
\end{aligned} \tag{22}$$

Thus, by choosing $c_0 = 0.5$, we have $\mathbb{P}\left(\Delta_a \geq -0.5, \forall a \in [L-1]\right) \geq 1 - L \exp\left(-\frac{m}{128\tau^2}\right)$. On the event $\{\Delta_a \geq -0.5, \forall a \in [L-1]\}$, we can use the relation $\log(1+x) \geq x - x^2$ for $x \geq -0.5$ and have

$$equation 19 \geq \log \|h_{i,0}\|^2 + \sum_{a=1}^{l} \left(\Delta_a - \Delta_a^2\right). \tag{23}$$

Due to (13) and (17), we have for any $c_1 > 0$, and $\tau^2 L \leq c_1/4$,

$$\mathbb{P}\left(\sum_{l=a}^{b}\xi_l \geq \frac{c_1}{2}\right) \leq \exp\left(-\frac{mc_1^2}{64\tau^2 L}\right),\ \mathbb{P}\left(\sum_{l=a}^{b}\xi_l < -\frac{c_1}{2}\right) \leq \exp\left(-\frac{mc_1^2}{64\tau^2 L}\right),$$
$$\mathbb{P}\left(\sum_{l=a}^{b}\zeta_l \geq \frac{c_1}{2}\right) \leq \exp\left(-\frac{mc_1^2}{4\tau^2 L}\right). \tag{24}$$

Thus we have for any $c_1 > 0$, and $\tau^2 L \leq c_1/4$,

$$\begin{aligned}\mathbb{P}\left(\sum_{a=1}^{l}\Delta_a \leq -c_1\right) &= \mathbb{P}\left(\sum_{a=1}^{l}\Delta_a \leq -c_1, \sum_{a=1}^{l}\xi_a \geq -\frac{c_1}{2}\right) + \mathbb{P}\left(\sum_{a=1}^{l}\Delta_a \leq -c_1, \sum_{a=1}^{l}\xi_a \leq -\frac{c_1}{2}\right)\\ &\leq \mathbb{P}\left(\sum_{a=1}^{l}\zeta_a \geq \frac{c_1}{2}\right) + \mathbb{P}\left(\sum_{a=1}^{l}\xi_a \leq -\frac{c_1}{2}\right) = 2\exp\left(-\frac{mc_1^2}{64\tau^2 L}\right).\end{aligned} \tag{25}$$

We can derive a similar result that $\mathbb{P}\left(\sum_{a=1}^{l}\Delta_a \geq c_1\right) \leq \mathbb{P}\left(\sum_{a=1}^{l}\xi_a \geq c_1\right) \leq \exp\left(-\frac{mc_1^2}{16\tau^2 L}\right)$. Let $a = b$ in (24), we have obtained that for a single Δ_a, for a constant c_1 such that $4\tau^2 \leq c_1$,

$$\mathbb{P}\left(|\Delta_a| \geq c_1\right) \leq 2\exp\left(-\frac{mc_1^2}{32\tau^2}\right). \tag{26}$$

In addition, we see that for any $16\tau^4 L \leq c_1$

$$\mathbb{P}\left(\sum_{a=1}^{l}\Delta_a^2 \geq c_1\right) \leq \sum_{a=1}^{l}\mathbb{P}\left(\Delta_a^2 \geq \frac{c_1}{l}\right) = \sum_{a=1}^{l}\mathbb{P}\left(|\Delta_a| \geq \sqrt{\frac{c_1}{l}}\right) \leq 2l\exp\left(-\frac{mc_1}{32}\right). \tag{27}$$

Thus, similar to the (25), we obtain for any $c_1 > 0$ and $8\tau^2 L < c_1$,

$$\begin{aligned}\mathbb{P}\left(\sum_{a=1}^{l}\left(\Delta_a - \Delta_a^2\right) \leq -c_1\right) &\leq \mathbb{P}\left(\sum_{a=1}^{l}\Delta_a \leq -\frac{c_1}{2}\right) + \mathbb{P}\left(\sum_{a=1}^{l}\Delta_a^2 \geq \frac{c_1}{2}\right)\\ &\leq 2\exp\left(-\frac{mc_1^2}{256\tau^2 L}\right) + 2(L-1)\exp\left(-\frac{mc_1}{64}\right)\\ &\leq 2L\exp\left(-\frac{mc_1}{64}\right)\end{aligned} \tag{28}$$

Thus on the event of $\{\Delta_a \geq -0.5, \forall a \in [L-1]\}$, we have for any $c_1 > 0$ and $8\tau^2 L < c_1$,

$$\mathbb{P}\left(\log\|h_{i,l}\|^2 \leq -c_1\right) \leq \mathbb{P}\left(\log\|h_{i,0}\|^2 + \sum_{a=1}^{l}\left(\Delta_a - \Delta_a^2\right) \leq -c_1\right) \leq 2L\exp\left(-\frac{mc_1}{64}\right). \tag{29}$$

Then we get the conclusion $\mathbb{P}\left(\|h_{i,l}\| < 1-c\right) = \mathbb{P}\left(\log\|h_{i,l}\|^2 \leq -2\log(1-c)^{-1}\right) \leq 2L\exp\left(-\frac{1}{32}m\log(1-c)^{-1}\right)$. Taking union bound over $i \in [n]$ and $l \in [L-1]$, we get the claimed result with probability $1 - 2nL^2\exp\left(-\frac{1}{32}m\log(1-c)^{-1}\right)$ under the condition $\tau^2 L \leq \frac{1}{4}\log(1-c)^{-1}$. □

C.2 Lemmas and proofs after perturbation

We use $\overrightarrow{\boldsymbol{W}}^{(0)}$ to denote the weight matrices at initialization and use $\overrightarrow{\boldsymbol{W}}'$ to denote the perturbation matrices. Let $\overrightarrow{\boldsymbol{W}} = \overrightarrow{\boldsymbol{W}}^{(0)} + \overrightarrow{\boldsymbol{W}}'$. We define $h_{i,l}^{(0)} = \phi((\boldsymbol{I} + \tau \boldsymbol{W}_l^{(0)})h_{i,l-1}^{(0)})$ and $h_{i,l} = \phi((\boldsymbol{I} + \tau \boldsymbol{W}_l)h_{i,l-1})$ for $l \in [L-1]$, and $h_{i,L}^{(0)} = \phi(\boldsymbol{W}_L^{(0)} h_{i,L-1}^{(0)})$ and $h_{i,L} = \phi(\boldsymbol{W}_L h_{i,L-1})$. Furthermore, let $h'_{i,l} := h_{i,l} - h_{i,l}^{(0)}$ and $\boldsymbol{D}'_{i,l} := \boldsymbol{D}_{i,l} - \boldsymbol{D}_{i,l}^{(0)}$. We note that $\|\cdot\|_0$ is the number of nonzero entries in $\cdot$. In the sequel, we will use notation O and Ω to simplify the presentation. Then the spectral norm bound after perturbation is as follows.

Lemma 3 *Suppose that $\overrightarrow{\boldsymbol{W}}^{(0)}$, $\boldsymbol{A}$ are randomly generated as in the initialization step, and $\boldsymbol{W}'_1, \ldots, \boldsymbol{W}'_{L-1} \in \mathbb{R}^{m\times m}$ are perturbation matrices with $\|\boldsymbol{W}'_l\| < \tau\omega$ for all $l \in [L-1]$ for some $\omega < 1$. Suppose $\boldsymbol{D}_{i,0}, \ldots, \boldsymbol{D}_{i,L}$ are diagonal matrices representing the activation status of sample i. If $\tau^2 L \le O(1)$, then with probability at least $1 - 3nL^2 \cdot \exp(-\Omega(m))$ over the initialization randomness we have*

$$\|(\boldsymbol{I} + \tau \boldsymbol{W}_b^{(0)} + \tau \boldsymbol{W}'_b)\boldsymbol{D}_{i,b-1} \cdots \boldsymbol{D}_{i,a}(\boldsymbol{I} + \tau \boldsymbol{W}_a^{(0)} + \tau \boldsymbol{W}'_a)\| \le O(1). \tag{30}$$

Proof This proof is similar to the proof of Theorem 1. We first build the claim for one fixed sample $i \in [n]$ and drop the subscript i, for convenience. We will show for a vector h_{a-1} with $\|h_{a-1}\| = 1$, we have $\|h_b\| \le 1 + c$ with high probability, where

$$h_b = \boldsymbol{D}_b(\boldsymbol{I} + \tau \boldsymbol{W}_b^{(0)} + \tau \boldsymbol{W}'_b)\boldsymbol{D}_{b-1} \cdots \boldsymbol{D}_a(\boldsymbol{I} + \tau \boldsymbol{W}_a^{(0)} + \tau \boldsymbol{W}'_a)h_{a-1}. \tag{31}$$

Let $g_l = h_{l-1} + \tau \boldsymbol{W}_l^{(0)} h_{l-1} + \tau \boldsymbol{W}'_l h_{l-1}$ and $h_l = \boldsymbol{D}_l g_l$ for $l = \{a, ..., b\}$. Then we have $\|g_l\| \ge \|h_l\|$ due to the fact $\|\boldsymbol{D}_l\| \le 1$. Hence we have

$$\|h_b\|^2 = \frac{\|h_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|h_a\|^2}{\|h_{a-1}\|^2}\|h_{a-1}\|^2 \le \frac{\|g_b\|^2}{\|h_{b-1}\|^2} \cdots \frac{\|g_a\|^2}{\|h_{a-1}\|^2}\|h_{a-1}\|^2.$$

Taking logarithm at both side, we have

$$\log \|h_b\|^2 \le \sum_{l=a}^{b} \log \Delta_l, \qquad \text{where } \Delta_l := \frac{\|g_l\|^2}{\|h_{l-1}\|^2}. \tag{32}$$

If letting $\tilde{h}_{l-1} := \frac{h_{l-1}}{\|h_{l-1}\|}$, then we obtain that

$$\begin{aligned}\log \Delta_l &= \log\left(1 + 2\tau\left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l^{(0)}\tilde{h}_{l-1}\right\rangle + \tau^2\|\boldsymbol{W}_l^{(0)}\tilde{h}_{l-1}\|^2 + 2\tau\left\langle (\boldsymbol{I} + \tau \boldsymbol{W}_l^{(0)})\tilde{h}_{l-1}, \boldsymbol{W}'_l\tilde{h}_{l-1}\right\rangle + \tau^2\|\boldsymbol{W}'_l\tilde{h}_{l-1}\|^2\right)\\ &\le 2\tau\left\langle \tilde{h}_{l-1}, \boldsymbol{W}_l^{(0)}\tilde{h}_{l-1}\right\rangle + \tau^2\|\boldsymbol{W}_l^{(0)}\tilde{h}_{l-1}\|^2 + 2\tau\left\langle (\boldsymbol{I} + \tau \boldsymbol{W}_l^{(0)})\tilde{h}_{l-1}, \boldsymbol{W}'_l\tilde{h}_{l-1}\right\rangle + \tau^2\|\boldsymbol{W}'_l\tilde{h}_{l-1}\|^2,\end{aligned}$$

where the inequality is due to the fact $\log(1 + x) \le x$ for all $x > -1$. We can bound the sum over layers of the first two terms as in the proof of Theorem 1. Next we control the last two terms related with $\boldsymbol{W}'_l$, on a high probability event $\{\|\boldsymbol{W}_l^{(0)}\| \le 4, \text{ for all } l \in [L-1]\}$

$$\begin{aligned}&\sum_{l=a}^{b} 2\tau\left\langle (\boldsymbol{I} + \tau \boldsymbol{W}_l^{(0)})\tilde{h}_{l-1}, \boldsymbol{W}'_l\tilde{h}_{l-1}\right\rangle \le \sum_{l=a}^{b} 2\tau\|\boldsymbol{I} + \tau \boldsymbol{W}_l^{(0)}\|\|\boldsymbol{W}'_l\|\|\tilde{h}_{l-1}\|^2 \le \sum_{l=a}^{b} 2\tau^2\omega(1 + 4\tau),\\ &\sum_{l=a}^{b} \tau^2\|\boldsymbol{W}'_l\tilde{h}_{l-1}\|^2 \le \sum_{l=a}^{b} 2\tau^4\omega^2.\end{aligned} \tag{33}$$

Hence given $\tau^2 L \le c_1/4$ as in proof of Theorem 1 and ω being a small constant, the above two sum are well controlled. We can obtain a spectral norm bound as claimed. Here the theorem is built for one $\boldsymbol{W}_l'$. At the end of the whole proof, we will see the number of iterations is $\Omega(n^2)$. If we take union bound over all the $\boldsymbol{W}_l'$ s running into in the optimization trajectory, the overall probability is still as high as $1 - \Omega(n^3L^2)\exp(-\Omega(m))$. □

We also have small changes on the output vector of each layer after perturbation.

Lemma 4 *Suppose that $\omega \le O(1)$ and $\tau^2 L \le O(1)$. If $\|\boldsymbol{W}_L'\| \le \omega$ and $\|\boldsymbol{W}_l'\| \le \tau\omega$ for $l \in [L-1]$, then with probability at least $1 - \exp(-\Omega(m\omega^{\frac{2}{3}}))$, the following bounds on $h_{i,l}'$ and $\boldsymbol{D}_{i,l}'$ hold for all $i \in [n]$ and all $l \in [L-1]$,*

$$\|h_{i,l}'\| \le O(\tau^2 L\omega), \quad \|\boldsymbol{D}_{i,l}'\|_0 \le O\Big(m(\omega\tau L)^{\frac{2}{3}}\Big), \quad \|h_{i,L}'\| \le O(\omega), \quad \|\boldsymbol{D}_{i,L}'\|_0 \le O\Big(m\omega^{\frac{2}{3}}\Big).$$

Proof Fixing i and ignoring the subscript in i, by Claim 8.2 in Allen-Zhu et al. (2018), for $l \in [L-1]$, there exists $\boldsymbol{D}_l''$ such that $|(\boldsymbol{D}_l'')_{k,k}| \le 1$ and

$$\begin{aligned}
h_l' &= \boldsymbol{D}_l''\Big((\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)} + \tau\boldsymbol{W}_l')h_{l-1} - (\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)})h_{l-1}^{(0)}\Big) \\
&= \boldsymbol{D}_l''\Big((\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)} + \tau\boldsymbol{W}_l')h_{l-1}' + \tau\boldsymbol{W}_l' h_{l-1}^{(0)}\Big) \\
&= \boldsymbol{D}_l''(\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)} + \tau\boldsymbol{W}_l')\boldsymbol{D}_{l-1}''(\boldsymbol{I} + \tau\boldsymbol{W}_{l-1} + \tau\boldsymbol{W}_{l-1}')h_{l-2}' \\
&\quad + \tau\boldsymbol{D}_l''(\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)} + \tau\boldsymbol{W}_l')\boldsymbol{D}_{l-1}''\boldsymbol{W}_{l-1}' h_{l-2}^{(0)} + \tau\boldsymbol{D}_l''\boldsymbol{W}_l' h_{l-1}^{(0)} \\
&= \cdots \\
&= \sum_{a=1}^{l} \tau\boldsymbol{D}_l''(\boldsymbol{I} + \tau\boldsymbol{W}_l^{(0)} + \tau\boldsymbol{W}_l') \cdots \boldsymbol{D}_{a+1}''(\boldsymbol{I} + \tau\boldsymbol{W}_{a+1} + \tau\boldsymbol{W}_{a+1}')\boldsymbol{D}_a''\boldsymbol{W}_a' h_a^{(0)}.
\end{aligned} \tag{34}$$

We claim that

$$\|h_l'\| \le O(\tau^2 L\omega) \tag{35}$$

due to the fact $\|\boldsymbol{D}_l''\| \le 1$ and the assumption $\|\boldsymbol{W}_l'\| \le \tau\omega$ for $l \in [L-1]$. This implies that $\|h_{i,l}'\|, \|g_{i,l}'\| \le O(\tau^2 L\omega)$ for all $l \in [L-1]$ and for all i with probability at least $1 - O(nL)\cdot\exp(-\Omega(m))$. One step further, we have $\|h_L'\|, \|g_L'\| \le O(\omega)$.

As for the sparsity $\|\boldsymbol{D}_l'\|_0$, we have $\|\boldsymbol{D}_l'\|_0 \le O(m(\omega\tau L)^{\frac{2}{3}})$ for every $l = [L-1]$ and $\|\boldsymbol{D}_L'\|_0 \le O(m\omega^{\frac{2}{3}})$.

The argument is as follows (adapt from the Claim 5.3 in Allen-Zhu et al. (2018)).

We first study the case $l \in [L-1]$. We see that if $(\boldsymbol{D}_l')_{j,j} \ne 0$ one must have $|(g_l')_j| > |(g_l^{(0)})_j|$.

We note that $(g_l^{(0)})_j = (h_{l-1}^{(0)} + \tau\boldsymbol{W}_l^{(0)}h_{l-1}^{(0)})_j \sim \mathcal{N}\Big((h_{l-1}^{(0)})_j, \frac{\tau^2\|h_{l-1}^{(0)}\|^2}{m}\Big)$. Let $\xi \le \frac{1}{\sqrt{m}}$be a parameter to be chosen later. Let $S_1 \subseteq [m]$ be a index set satisfying $S_1 := \{j : |(g_l^{(0)})_j| \le \xi\tau\}$. We have $\mathbb{P}\{|(g_l^{(0)})_j| \le \xi\tau\} \le O(\xi\sqrt{m})$ for each $j \in [m]$. By Chernoff bound, with probability at least $1 - \exp(-\Omega(m^{3/2}\xi))$ we have

$$|S_1| \le O(\xi m^{3/2}).$$

Let $S_2 := \{j : j \notin S_1, \text{ and } (\boldsymbol{D}_l')_{j,j} \ne 0\}$. Then for $j \in S_2$, we have $|(g_l')_j| > \xi\tau$. As we have proved that $\|g_l'\| \le O(\tau^2 L\omega)$, we have

$$|S_2| \le \frac{\|g_l'\|^2}{(\xi\tau)^2} = O((\omega\tau L)^2/\xi^2).$$

Choosing ξ to minimize $|S_1| + |S_2|$, we have $\xi = (\omega\tau L)^{\frac{2}{3}}/\sqrt{m}$ and consequently, $\|\boldsymbol{D}_l'\|_0 \le O(m(\omega\tau L)^{\frac{2}{3}})$. Similarly, we have $\|\boldsymbol{D}_L'\|_0 \le O(m\omega^{\frac{2}{3}})$. □

We next prove that the norm of a sparse vector after the ResNet mapping.

Lemma 5 *Suppose that $s \ge \Omega(d/\log m)$, $\tau^2 L \le O(1)$. If $\boldsymbol{W}_l$ for $l \in [L]$ satisfy the condition as in Lemma 3, then for all $i \in [n]$ and $a \in [L]$ and for all s-sparse vectors $u \in \mathbb{R}^m$ and for all $v \in \mathbb{R}^d$, the following bound holds with probability at least $1 - (nL) \cdot \exp(-\Omega(s \log m))$*

$$|v^T \boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L\boldsymbol{D}_{i,L-1}(\boldsymbol{I} + \tau\boldsymbol{W}_{L-1}) \cdots \boldsymbol{D}_{i,a}(\boldsymbol{I} + \tau\boldsymbol{W}_a)u| \le O\left(\frac{\sqrt{s\log m}}{\sqrt{d}}\|u\|\|v\|\right), \quad (36)$$

where $\boldsymbol{D}_{i,a}$ is diagonal activation matrix for sample i.

Proof For any fixed vector $u \in \mathbb{R}^m$, $\|\boldsymbol{D}_{i,L}\boldsymbol{W}_L\boldsymbol{D}_{i,L-1}(\boldsymbol{I} + \tau\boldsymbol{W}_{L-1}) \cdots \boldsymbol{D}_{i,a}(\boldsymbol{I} + \tau\boldsymbol{W}_a)u\| \le 1.1\|u\|$ holds with probability at least $1 - \exp(-\Omega(m))$ because of Lemma 3.

On the above event, for a fixed vector $v \in \mathbb{R}^d$ and any fixed $\boldsymbol{W}_l$ for $l \in [L]$, the randomness only comes from $\boldsymbol{B}$, then $v^T\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L\boldsymbol{D}_{i,L-1}(\boldsymbol{I} + \tau\boldsymbol{W}_{L-1}) \cdots \boldsymbol{D}_{i,a}(\boldsymbol{I} + \tau\boldsymbol{W}_a)u$ is a Gaussian variable with mean 0 and variance no larger than $1.1^2\|u\|^2 \cdot \|v\|^2/d$. Hence

$$\begin{aligned}&\mathbb{P}\big\{|v^T\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L\boldsymbol{D}_{i,L-1}(\boldsymbol{I} + \tau\boldsymbol{W}_{L-1}) \cdots \boldsymbol{D}_{i,a}(\boldsymbol{I} + \tau\boldsymbol{W}_a)u| \ge \sqrt{s\log m} \cdot \Omega(\|u\|\|v\|/\sqrt{d})\big\}\\ &\quad = \mathrm{erfc}(\Omega(\sqrt{s\log m})) \le \exp(-\Omega(s\log m)).\end{aligned}$$

Take ϵ-net over all s-sparse vectors of u and all d-dimensional vectors of v, if $s \ge \Omega(d/\log m)$ then with probability $1 - \exp(-\Omega(s\log m))$ the claim holds for all s-sparse vectors of u and all d-dimensional vectors of v. Further taking the union bound over all $i \in [n]$ and $a \in [L]$, the lemma is proved. □

D Gradient lower/upper bounds and their proofs

Because the gradient is pathological and data-dependent, in order to build bound on the gradient, we need to consider all possible point and all cases of data. Hence we first introduce an arbitrary loss vector and then the gradient bound can be obtained by taking a union bound.

We define the $\mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(v, \cdot)$ operator. It back-propagates a vector v to the $\cdot$ which could be the intermediate output h_l or the parameter $\boldsymbol{W}_l$ at the specific layer l using the forward propagation state of input i through the network with parameter $\overrightarrow{\boldsymbol{W}}$. Specifically,

$$\begin{aligned}\mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(v, h_l) &:= (\boldsymbol{I} + \tau\boldsymbol{W}_{l+1})^T\boldsymbol{D}_{i,l+1} \cdots (\boldsymbol{I} + \tau\boldsymbol{W}_{L-1})^T\boldsymbol{D}_{i,L-1}\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\boldsymbol{B}^T v,\\ \mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(v, \boldsymbol{W}_l) &:= \tau\big(\boldsymbol{D}_{i,l}(\boldsymbol{I} + \tau\boldsymbol{W}_{l+1})^T \cdots (\boldsymbol{I} + \tau\boldsymbol{W}_{L-1})^T\boldsymbol{D}_{i,L-1}\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\boldsymbol{B}^T v\big)h_{i,l-1}^T \quad \forall l \in [L-1],\\ \mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(v, \boldsymbol{W}_L) &:= \big(\boldsymbol{D}_{i,L}\boldsymbol{B}^T v\big)h_{i,L-1}^T.\end{aligned}$$

Moreover, we introduce

$$\mathsf{BP}_{\overrightarrow{\boldsymbol{W}}}(\vec{v}, \boldsymbol{W}_l) := \sum_{i=1}^{n} \mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(v_i, \boldsymbol{W}_l) \quad \forall l \in [L],$$

where $\vec{v}$ is composed of n vectors v_i for $i \in [n]$. If v_i is the error signal of input i, then $\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}}) = \mathsf{BP}_{\overrightarrow{\boldsymbol{W}},i}(\boldsymbol{B}h_{i,L} - y_i^*, \boldsymbol{W}_l)$.

D.1 Gradient upper bound

Proof We ignore the superscript $^{(0)}$ for simplicity. Then for an $i \in [n]$ we have

$$\left\|\nabla_{\boldsymbol{W}_L} F_i(\overrightarrow{\boldsymbol{W}})\right\|_F = \left\|\left(\boldsymbol{D}_{i,L}\partial h_{i,L}\right)h_{i,L-1}^T\right\|_F = \left\|\left(\boldsymbol{D}_{i,L}\partial h_{i,L}\right)\right\|\left\|h_{i,L-1}^T\right\| \le \frac{1+c}{1-\epsilon}\|\partial h_{i,L}\|,$$

because of Theorem 1. Similarly, we have for $l \in [L-1]$,

$$\begin{aligned}\left\|\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}})\right\|_F &= \left\|\tau\left(\boldsymbol{D}_{i,l}(\boldsymbol{I}+\tau\boldsymbol{W}_{l+1})^T\cdots(\boldsymbol{I}+\tau\boldsymbol{W}_{L-1})^T\boldsymbol{D}_{i,L-1}\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\partial h_{i,L}\right)h_{i,l-1}^T\right\|_F \\ &\le \tau\|\boldsymbol{D}_{i,l}(\boldsymbol{I}+\tau\boldsymbol{W}_{l+1})^T\cdots\boldsymbol{D}_{i,L-1}\|\cdot\|\boldsymbol{W}_L^T\boldsymbol{D}_{i,L}\|\cdot\|\partial h_{i,L}\|\cdot\|h_{i,l-1}\| \\ &\le \frac{(1+c)^2}{(1-\epsilon)^2}(2\sqrt{2}+c)\tau\|\partial h_{i,L}\|,\end{aligned}$$

because of Theorem 1 and Lemma 2. □

The above upper bounds hold for the initialization $\overrightarrow{\boldsymbol{W}}^{(0)}$ because of Theorem 1 and Theorem 2. They also hold for all the $\overrightarrow{\boldsymbol{W}}$ such that $\|\overrightarrow{\boldsymbol{W}} - \overrightarrow{\boldsymbol{W}}^{(0)}\| \le \omega$ due to Lemma 3.

For the quadratic loss function, we have $\|\partial h_{i,L}\|^2 = \|\boldsymbol{B}^T(\boldsymbol{B}h_{i,L} - y_i^*)\|^2 = O(m/d)F_i(\overrightarrow{\boldsymbol{W}})$. We have the gradient upper bound as follows.

Theorem 6 *Suppose $\omega = O(1)$. For every input sample $i \in [n]$ and for every $l \in [L-1]$ and for every $\overrightarrow{\boldsymbol{W}}$ such that $\|\boldsymbol{W}_L - \boldsymbol{W}_L^{(0)}\| \le \omega$ and $\|\boldsymbol{W}_l - \boldsymbol{W}_l^{(0)}\| \le \tau\omega$, the following holds with probability at least $1 - O(nL^2)\cdot\exp(-\Omega(m))$ over the randomness of $\boldsymbol{A}, \boldsymbol{B}$ and $\overrightarrow{\boldsymbol{W}}^{(0)}$*

$$\begin{aligned}\|\nabla_{\boldsymbol{W}_l} F_i(\overrightarrow{\boldsymbol{W}})\|_F^2 &\le O\left(\frac{\tau^2 m}{d}F_i(\overrightarrow{\boldsymbol{W}})\right), \\ \|\nabla_{\boldsymbol{W}_L} F_i(\overrightarrow{\boldsymbol{W}})\|_F^2 &\le O\left(\frac{m}{d}F_i(\overrightarrow{\boldsymbol{W}})\right).\end{aligned} \tag{37}$$

D.2 Gradient lower bound

For the quadratic loss function, we have the following gradient lower bound.

Theorem 7 *Let $\omega = O\left(\frac{\delta^{3/2}}{n^3\log^3 m}\right)$. With probability at least $1 - \exp(-\Omega(m\omega^{\frac{2}{3}}))$ over the randomness of $\overrightarrow{\boldsymbol{W}}^{(0)}, \boldsymbol{A}, \boldsymbol{B}$, it satisfies for every $\overrightarrow{\boldsymbol{W}}$ with $\|\overrightarrow{\boldsymbol{W}} - \overrightarrow{\boldsymbol{W}}^{(0)}\| \le \omega$,*

$$\|\nabla_{\boldsymbol{W}_L} F(\overrightarrow{\boldsymbol{W}})\|_F^2 \geq \Omega\left(\frac{F(\overrightarrow{\boldsymbol{W}})}{dn/\delta} \times m\right). \tag{38}$$

This gradient lower bound on $\|\nabla_{\boldsymbol{W}_L} F(\overrightarrow{\boldsymbol{W}})\|_F^2$ acts like the gradient dominance condition (Zou and Gu, 2019; Allen-Zhu et al., 2018) except that our range on ω does not depend on the depth L.

Proof The gradient lower-bound at the initialization is given by the Section 6.2 in (Allen-Zhu et al., 2018) and the Lemma 4.1 in (Zou and Gu, 2019) via the smoothed analysis (Spielman and Teng, 2004): with high probability the gradient is lower-bounded, although the worst case it might be 0. We adopt the same proof for the Lemma 4.1 in Zou and Gu (2019) based on two preconditioned results Theorem 2 and Lemma 6. We shall not repeat it here.

Now suppose that we have $\|\nabla_{\boldsymbol{W}_L} F(\overrightarrow{\boldsymbol{W}}^{(0)})\|_F^2 \geq \Omega\left(\frac{F(\overrightarrow{\boldsymbol{W}}^{(0)})}{dn/\delta} \times m\right)$. We next bound the change of the gradient after perturbing the parameter. Recall that

$$\mathsf{BP}_{\overrightarrow{\boldsymbol{W}}^{(0)}}(\vec{v}, \boldsymbol{W}_L) - \mathsf{BP}_{\overrightarrow{\boldsymbol{W}}}(\vec{v}, \boldsymbol{W}_L) = \sum_{i=1}^{n} \left((v_i^T \boldsymbol{BD}_{i,L}^{(0)})^T (h_{i,L-1}^{(0)})^T - (v_i^T \boldsymbol{BD}_{i,L})^T (h_{i,L-1})^T \right)$$

By Lemma 4 and Lemma 5, we know,

$$\|v_i^T \boldsymbol{BD}_{i,L}^{(0)} - v_i^T \boldsymbol{BD}_{i,L}\| \leq O(\sqrt{m\omega^{\frac{2}{3}}}/\sqrt{d}) \cdot \|v_i\|.$$

Furthermore, we know

$$\|v_i^T \boldsymbol{BD}_{i,L}\| \leq O(\sqrt{m/d}) \cdot \|v_i\|.$$

By Theorem 2 and Lemma 4, we have

$$\|h_{i,L-1}^{(0)}\| \leq 1.1 \quad \text{and} \quad \|h_{i,L-1} - h_{i,L-1}^{(0)}\| \leq O(\omega).$$

Combing the above bounds together, we have

$$\|\mathsf{BP}_{\overrightarrow{\boldsymbol{W}}^{(0)}}(\vec{v}, \boldsymbol{W}_L) - \mathsf{BP}_{\overrightarrow{\boldsymbol{W}}}(\vec{v}, \boldsymbol{W}_L)\|_F^2 \leq n\|\vec{v}\|^2 \cdot O(\sqrt{m\omega^{\frac{2}{3}}/d} + \omega\sqrt{m/d})^2 \leq n\|\vec{v}\|^2 \cdot O\left(\frac{m}{d}\omega^{\frac{2}{3}}\right)$$

Hence the gradient lower bound still holds for $\overrightarrow{\boldsymbol{W}}$ given $\omega < O\left(\frac{\delta^{3/2}}{n^3}\right)$.

Finally, taking $\epsilon-$net over all possible vectors $\vec{v} = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$, we prove that the above gradient lower bound holds for all $\vec{v}$. In particular, we can now plug in the choice of $v_i = \boldsymbol{B}h_{i,L} - y_i^*$ and it implies our desired bounds on the true gradients. □

The gradient lower bound requires the following property.

Lemma 6 *For any δ and any pair (x_i, x_j) satisfying $\|x_i - x_j\| \geq \delta$, then $\|h_{j,l} - h_{j,l}\| \geq \Omega(\delta)$ holds for all $l \in [L]$ with probability at least $1 - O(n^2 L) \cdot \exp(-\Omega(\log^2 m))$ given that $\tau \leq O(1/(\sqrt{L}\log m))$ and $m \geq \Omega(\tau^2 L^2 \delta^{-2})$.*

The proof of Lemma 6 follows the Appendix C in Allen-Zhu et al. (2018).

E Semi-smoothness for $\tau \leq O(1/\sqrt{L})$

With the help of Theorem 6 and several other improvements, we can obtain a tighter bound on the semi-smoothness condition of the objective function.

Theorem 8 *Let* $\omega = O\left(\frac{\delta^{3/2}}{n^3L^{7/2}}\right)$ *and* $\tau^2 L \leq 1$. *With high probability, we have for every* $\breve{\overrightarrow{\boldsymbol{W}}} \in (\mathbb{R}^{m\times m})^L$ *with* $\left\|\breve{\overrightarrow{\boldsymbol{W}}} - \overrightarrow{\boldsymbol{W}}^{(0)}\right\| \leq \omega$ *and for every* $\overrightarrow{\boldsymbol{W}}' \in (\mathbb{R}^{m\times m})^L$ *with* $\|\overrightarrow{\boldsymbol{W}}'\| \leq \omega$, *we have*

$$F(\breve{\overrightarrow{\boldsymbol{W}}} + \overrightarrow{\boldsymbol{W}}') \leq F(\breve{\overrightarrow{\boldsymbol{W}}}) + \langle \nabla F(\breve{\overrightarrow{\boldsymbol{W}}}), \overrightarrow{\boldsymbol{W}}'\rangle + O(\frac{nm}{d})\|\overrightarrow{\boldsymbol{W}}'\|_F^2 + O\left(\sqrt{\frac{m}{nd}}\omega^{\frac{1}{3}}L^{\frac{7}{6}}\right)\|\overrightarrow{\boldsymbol{W}}'\|_F\sqrt{F(\breve{\overrightarrow{\boldsymbol{W}}})}.$$

We will show the semi-smoothness theorem for a more general $\omega \in \left[\Omega\left((d/(m\log m))^{\frac{3}{2}}\right), O(1)\right]$ and the above high probability is at least $1 - \exp(-\Omega(m\omega^{\frac{2}{3}}))$ over the randomness of $\overrightarrow{\boldsymbol{W}}^{(0)}, \boldsymbol{A}, \boldsymbol{B}$.

Before going to the proof of the theorem, we introduce a lemma.

Lemma 7 *There exist diagonal matrices* $\boldsymbol{D}''_{i,l} \in \mathbb{R}^{m\times m}$ *with entries in* [-1,1] *such that* $\forall i \in [n], \forall l \in [L-1]$,

$$h_{i,l} - \breve{h}_{i,l} = \sum_{a=1}^{l}(\breve{\boldsymbol{D}}_{i,l} + \boldsymbol{D}''_{i,l})(\boldsymbol{I} + \tau\breve{\boldsymbol{W}}_l)\cdots(\boldsymbol{I} + \tau\breve{\boldsymbol{W}}_{a+1})(\breve{\boldsymbol{D}}_{i,a} + \boldsymbol{D}''_{i,a})\tau\boldsymbol{W}'_a h_{i,a-1}, \tag{39}$$

and

$$\begin{aligned} h_{i,L} - \breve{h}_{i,L} =& (\breve{\boldsymbol{D}}_{i,L} + \boldsymbol{D}''_{i,L})\boldsymbol{W}'_L h_{i,L-1} \\ &+ \sum_{a=1}^{L-1}(\breve{\boldsymbol{D}}_{i,L} + \boldsymbol{D}''_{i,L})\breve{\boldsymbol{W}}_L\cdots(\boldsymbol{I} + \tau\breve{\boldsymbol{W}}_{a+1})(\breve{\boldsymbol{D}}_{i,a} + \boldsymbol{D}''_{i,a})\tau\boldsymbol{W}'_a h_{i,a-1}. \end{aligned} \tag{40}$$

Furthermore, we then have $\forall l \in [L-1], \|h_{i,l} - \breve{h}_{i,l}\| \leq O(\tau^2 L\omega)$, $\|\boldsymbol{D}''_{i,l}\|_0 \leq O(m(\omega\tau L)^{\frac{2}{3}})$, *and* $\|h_{i,L} - \breve{h}_{i,L}\| \leq O((1+\tau\sqrt{L})\|\boldsymbol{W}'\|_F)$, $\|\boldsymbol{D}''_{i,L}\|_0 \leq O(m\omega^{\frac{2}{3}})$ *and*

$$\|\boldsymbol{B}h_{i,L} - \boldsymbol{B}\breve{h}_{i,L}\| \leq O(\sqrt{m/d})\|\boldsymbol{W}'\|_F$$

hold with probability $1 - \exp(-\Omega(m\omega^{\frac{2}{3}}))$ *given* $\|\boldsymbol{W}'_L\| \leq \omega, \|\boldsymbol{W}'_l\| \leq \tau\omega$ *for* $l \in [L-1]$ *and* $\omega \leq O(1), \tau\sqrt{L} \leq 1$.

Proof of Theorem 8 First of all, we know that $\breve{loss}_i := \boldsymbol{B}\breve{h}_{i,L} - y_i^*$

$$\begin{aligned} \frac{1}{2}\|\boldsymbol{B}h_{i,L} - y_i^*\|^2 &= \frac{1}{2}\|\breve{loss}_i + \boldsymbol{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \\ &= \frac{1}{2}\|\breve{loss}_i\|^2 + \breve{loss}_i^T\boldsymbol{B}(h_{i,L} - \breve{h}_{i,L}) + \frac{1}{2}\|\boldsymbol{B}(h_{i,L} - \breve{h}_{i,L})\|^2, \end{aligned} \tag{41}$$

and

$$\begin{aligned}\nabla_{\boldsymbol{W}_l} F(\overrightarrow{\boldsymbol{W}}) &= \sum_{i=1}^{n} (loss_i^T \boldsymbol{B} \boldsymbol{D}_{i,L} \boldsymbol{W}_L \cdots \boldsymbol{D}_{i,l+1} (\boldsymbol{I} + \tau \boldsymbol{W}_{l+1}) \boldsymbol{D}_{i,l})^T (\tau h_{i,l-1})^T. \\ \nabla_{\boldsymbol{W}_L} F(\overrightarrow{\boldsymbol{W}}) &= \sum_{i=1}^{n} (loss_i^T \boldsymbol{B} \boldsymbol{D}_{i,L})^T (h_{i,l-1})^T.\end{aligned} \tag{42}$$

We use the relation that for two matrices A, B, $\langle A, B \rangle = \mathrm{tr}(A^T B)$. Then, we can write

$$\langle \nabla_{\boldsymbol{W}_l} F(\overrightarrow{\breve{\boldsymbol{W}}}), \boldsymbol{W}_l' \rangle = \sum_{i=1}^{n} (\breve{loss}_i^T \boldsymbol{B} \breve{\boldsymbol{D}}_{i,L} \breve{\boldsymbol{W}}_L \cdots (\boldsymbol{I} + \tau \breve{\boldsymbol{W}}_{l+1}) \breve{\boldsymbol{D}}_{i,l} \boldsymbol{W}_l' (\tau \breve{h}_{i,l-1}). \tag{43}$$

Then further by Lemma 7, we have

$$\begin{aligned}
&F(\overrightarrow{\breve{\boldsymbol{W}}} + \overrightarrow{\boldsymbol{W}'}) - F(\overrightarrow{\breve{\boldsymbol{W}}}) - \langle \nabla F(\overrightarrow{\breve{\boldsymbol{W}}}), \overrightarrow{\boldsymbol{W}'} \rangle \\
&= -\langle \nabla F(\overrightarrow{\breve{\boldsymbol{W}}}), \overrightarrow{\boldsymbol{W}'} \rangle + \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{B} h_{i,L} - y_i^*\|^2 - \|\boldsymbol{B} \breve{h}_{i,L} - y_i^*\|^2 \\
&= -\sum_{l=1}^{L} \langle \nabla_{\boldsymbol{W}_l} F(\overrightarrow{\breve{\boldsymbol{W}}}), \boldsymbol{W}_l' \rangle + \sum_{i=1}^{n} \breve{loss}_i^T \boldsymbol{B} (h_{i,L} - \breve{h}_{i,L}) + \frac{1}{2} \|\boldsymbol{B} (h_{i,L} - \breve{h}_{i,L})\|^2 \\
&\overset{(a)}{=} \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{B} (h_{i,L} - \breve{h}_{i,L})\|^2 + \sum_{i=1}^{n} \breve{loss}_i^T \boldsymbol{B} \Big((\breve{\boldsymbol{D}}_{i,L} + \boldsymbol{D}_{i,L}'') \boldsymbol{W}_L' h_{i,L-1} - (\breve{\boldsymbol{D}}_{i,L}) \boldsymbol{W}_L' \breve{h}_{i,L-1} \Big) \\
&\quad + \sum_{i=1}^{n} \sum_{l=1}^{L-1} \breve{loss}_i^T \boldsymbol{B} \Big((\breve{\boldsymbol{D}}_{i,L} + \boldsymbol{D}_{i,L}'') \breve{\boldsymbol{W}}_L \cdots (\boldsymbol{I} + \tau \breve{\boldsymbol{W}}_{l+1}) (\breve{\boldsymbol{D}}_{i,l} + \boldsymbol{D}_{i,l}'') \tau \boldsymbol{W}_l' h_{i,l-1} \\
&\quad - \breve{\boldsymbol{D}}_{i,L} \breve{\boldsymbol{W}}_L \cdots (\boldsymbol{I} + \tau \breve{\boldsymbol{W}}_{l+1}) \breve{\boldsymbol{D}}_{i,l} \boldsymbol{W}_l' (\tau \breve{h}_{i,l-1}) \Big),
\end{aligned} \tag{44}$$

where (a) is due to Lemma 7.

We next bound the RHS of (45). We first use Lemma 7 to get

$$\|\boldsymbol{B} (h_{i,L} - \breve{h}_{i,L})\| \leq O(\sqrt{m/d}) \|\boldsymbol{W}'\|_F. \tag{45}$$

Next we calculate that for $l = L$,

$$\begin{aligned}
&\Big| \breve{loss}_i^T \boldsymbol{B} \Big((\breve{\boldsymbol{D}}_{i,L} + \boldsymbol{D}_{i,L}'') \boldsymbol{W}_L' h_{i,L-1} - (\breve{\boldsymbol{D}}_{i,L}) \boldsymbol{W}_L' \breve{h}_{i,L-1} \Big) \Big| \\
&\leq \Big| \breve{loss}_i^T \boldsymbol{B} \Big(\boldsymbol{D}_{i,L}'' \boldsymbol{W}_L' h_{i,L-1} \Big) \Big| + \Big| \breve{loss}_i^T \boldsymbol{B} \Big(\breve{\boldsymbol{D}}_{i,L} \boldsymbol{W}_L' (h_{i,L-1} - \breve{h}_{i,L-1}) \Big) \Big|.
\end{aligned} \tag{46}$$

For the first term, by Lemma 5 and Lemma 7, we have

$$\begin{aligned}\left|\breve{loss}_i^T \boldsymbol{B}\left(\boldsymbol{D}''_{i,L}\boldsymbol{W}'_L h_{i,L-1}\right)\right| &\le O\left(\frac{\sqrt{m\omega^{\frac{2}{3}}}}{\sqrt{d}}\right)\|\breve{loss}_i\|\cdot\|\boldsymbol{W}'_L h_{i,L-1}\| \\ &\le O\left(\frac{\sqrt{m\omega^{\frac{2}{3}}}}{\sqrt{d}}\right)\|\breve{loss}_i\|\cdot\|\boldsymbol{W}'_L\|,\end{aligned} \tag{47}$$

where the last inequality is due to $\|h_{i,L-1}\| \le O(1)$. For the second term, by Lemma 7 we have

$$\begin{aligned}&\left|\breve{loss}_i^T \boldsymbol{B}\left(\breve{\boldsymbol{D}}_{i,L}\boldsymbol{W}'_L(h_{i,L-1}-\breve{h}_{i,L-1})\right)\right| \\ &\quad\le \|\breve{loss}_i\|\cdot\left\|\boldsymbol{B}\breve{\boldsymbol{D}}_{i,L}\right\|_2\cdot\|\boldsymbol{W}'_L\|\|h_{i,L-1}-\breve{h}_{i,L-1}\| \\ &\quad\le \|\breve{loss}_i\|\cdot O\left(\frac{\omega\sqrt{m}}{\sqrt{d}}\right)\cdot\|\boldsymbol{W}'_L\|,\end{aligned} \tag{48}$$

where the last inequality is due to the assumption $\|\boldsymbol{W}'_L\| \le \omega$. Similarly for $l \in [L-1]$, we ignore the index i for simplicity.

$$\begin{aligned}&\Big|\sum_{l=1}^{L-1}\breve{loss}^T\Big(\boldsymbol{B}(\breve{\boldsymbol{D}}_L+\boldsymbol{D}''_L)\breve{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{l+1})(\breve{\boldsymbol{D}}_l+\boldsymbol{D}''_l)-\boldsymbol{B}\breve{\boldsymbol{D}}_L\breve{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{l+1})\breve{\boldsymbol{D}}_l\Big)\boldsymbol{W}'_l(\tau h_{l-1})\Big| \\ &\quad=\Big|\sum_{l=1}^{L-1}\breve{loss}^T\boldsymbol{B}\boldsymbol{D}''_L\breve{\boldsymbol{W}}_L(\boldsymbol{D}_{L-1}+\boldsymbol{D}''_{L-1})(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{L-1})\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)(\tau\boldsymbol{W}'_l h_{l-1})\Big| \\ &\qquad+\Big|\sum_{l=1}^{L-1}\sum_{a=l}^{L-1}\breve{loss}^T\boldsymbol{B}\breve{\boldsymbol{D}}_L\breve{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{a+1})\boldsymbol{D}''_a(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_a)\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)(\tau\boldsymbol{W}'_l h_{l-1})\Big| \\ &\qquad+\Big|\sum_{l=1}^{L-1}\breve{loss}^T\boldsymbol{B}\breve{\boldsymbol{D}}_L\breve{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{l+1})\breve{\boldsymbol{D}}_l\boldsymbol{W}'_l\tau(h_{l-1}-\breve{h}_{l-1})\Big|\end{aligned} \tag{49}$$

We next bound the terms in (50) one by one. For the first term, by Lemma 5 and Lemma 7, we have

$$\begin{aligned}&\left|\sum_{l=1}^{L-1}\breve{loss}^T\boldsymbol{B}\boldsymbol{D}''_L\breve{\boldsymbol{W}}_L(\boldsymbol{D}_{L-1}+\boldsymbol{D}''_{L-1})(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{L-1})\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)(\tau\boldsymbol{W}'_l h_{l-1})\right| \\ &\quad\le O\left(\frac{\sqrt{m\omega^{\frac{2}{3}}}}{\sqrt{d}}\right)\left\|\breve{loss}\right\|\cdot\left\|\sum_{l=1}^{L-1}\breve{\boldsymbol{W}}_L(\boldsymbol{D}_{L-1}+\boldsymbol{D}''_{L-1})(\boldsymbol{I}+\tau\breve{\boldsymbol{W}}_{L-1})\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)(\tau\boldsymbol{W}'_l h_{l-1})\right\| \\ &\quad\overset{(a)}{\le} O\left(\frac{\sqrt{m\omega^{\frac{2}{3}}}}{\sqrt{d}}\right)\cdot\|\breve{loss}\|\cdot\tau\sqrt{L}\|\boldsymbol{W}'_{L-1:1}\|_F,\end{aligned} \tag{50}$$

where $\|\boldsymbol{W}'_{L-1:1}\|_F = \sqrt{\sum_{l=1}^{L-1}\|\boldsymbol{W}'_l\|_F^2}$ and (a) is due to the similar argument (56) in the proof Lemma 7 and the fact $\left\|\check{\boldsymbol{W}}_L(\boldsymbol{D}_{L-1}+\boldsymbol{D}''_{L-1})(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{L-1})\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)\right\| = O(1)$ and $\|h_{L-1}\| = O(1)$ holds with high probability. We note that the inequality (a) helps us save a $\sqrt{L}$ factor in our main theorem.

We have similar bound for the second term of (50)

$$\begin{aligned}
&\left|\sum_{l=1}^{L-1}\sum_{a=l}^{L-1}\check{loss}^T\boldsymbol{B}\check{\boldsymbol{D}}_L\check{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{a+1})\boldsymbol{D}''_a(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_a)\cdots(\boldsymbol{D}_l+\boldsymbol{D}''_l)(\tau\boldsymbol{W}'_l h_{l-1})\right| \\
&\quad\le O\left(\frac{\sqrt{m(\omega\tau L)^{\frac{2}{3}}}}{\sqrt{d}}\right)\cdot\|\check{loss}\|\cdot\tau\sum_{a=1}^{L-1}\sqrt{a}\|\boldsymbol{W}'_{a:1}\|_F \\
&\quad\le O\left(\frac{\sqrt{m(\omega\tau L)^{\frac{2}{3}}}}{\sqrt{d}}\right)\cdot\|\check{loss}\|\cdot\tau L^{3/2}\|\boldsymbol{W}'_{L-1:1}\|_F.
\end{aligned} \tag{51}$$

For the last term in (50), we have

$$\begin{aligned}
&\left|\sum_{l=1}^{L-1}\check{loss}^T\boldsymbol{B}\check{\boldsymbol{D}}_L\check{\boldsymbol{W}}_L\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{l+1})\check{\boldsymbol{D}}_l\boldsymbol{W}'_l\tau(h_{l-1}-\check{h}_{l-1})\right| \\
&\quad\le\|\check{loss}\|\cdot O\left(\sqrt{m/d}\right)\cdot\sum_{l=1}^{L-1}\|\boldsymbol{W}'_l\|\cdot\tau^3 L\omega \\
&\quad\le\|\check{loss}\|\cdot O\left(\sqrt{m/d}\right)\cdot\|\boldsymbol{W}'_{L-1:1}\|_F\cdot(\tau^2 L)^{3/2},
\end{aligned} \tag{52}$$

where is the last inequality is due to the bound on $\|h_{l-1}-\check{h}_{l-1}\|$ in Lemma 7. Hence

$$\begin{aligned}
equation50 &\le O\left(\frac{\sqrt{m(\omega\tau L)^{\frac{2}{3}}}}{\sqrt{d}}\right)\cdot\|\check{loss}\|\cdot\tau L^{3/2}\|\boldsymbol{W}_{L-1:1}\|_F \\
&\le O\left((\tau L)^{\frac{4}{3}}\frac{\sqrt{mL\omega^{\frac{2}{3}}}}{\sqrt{d}}\right)\cdot\|\check{loss}\|\cdot\|\boldsymbol{W}'_{L-1:1}\|_F.
\end{aligned} \tag{53}$$

Having all the above together and using triangle inequality, we have the result. □

Proof of Lemma 7 The proof relies on the following lemma.

Lemma 8 (Proposition 8.3 in in Allen-Zhu et al. (2018)) *Given vectors $a, b \in \mathbb{R}^m$ and $\boldsymbol{D} \in \mathbb{R}^{m\times m}$ the diagonal matrix where $\boldsymbol{D}_{k,k} = \mathbf{1}_{a_k\ge 0}$. Then, there exists a diagonal matrix $\boldsymbol{D}'' \in \mathbb{R}^{m\times m}$ with*

- $|\boldsymbol{D}_{k,k}+\boldsymbol{D}''_{k,k}|\le 1$ and $|\boldsymbol{D}''_{k,k}|\le 1$ for every $k\in[m]$,
- $\boldsymbol{D}''_{k,k}\neq 0$ only when $\mathbf{1}_{a_k\ge 0}\neq \mathbf{1}_{b_k\ge 0}$,
- $\phi(a)-\phi(b)=(\boldsymbol{D}+\boldsymbol{D}'')(a-b)$.

Fixing index i and ignoring the subscript in i for simplicity, by Lemma 8, for each $l\in[L-1]$ there exists a $\boldsymbol{D}''_l$ such that $|(\boldsymbol{D}''_l)_{k,k}|\le 1$ and

$$\begin{aligned} h_l-\check{h}_l &= \phi((\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l+\tau\boldsymbol{W}'_l)h_{l-1})-\phi((\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\check{h}_{l-1})\\ &= (\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)\Big((\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l+\tau\boldsymbol{W}'_l)h_{l-1}-(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\check{h}_{l-1}\Big)\\ &= (\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)(h_{l-1}-\check{h}_{l-1})+(\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)\tau\boldsymbol{W}'_l h_{l-1}\\ &= \sum_{a=1}^{l}(\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{a+1})(\check{\boldsymbol{D}}_a+\boldsymbol{D}''_a)\tau\boldsymbol{W}'_a h_{a-1} \end{aligned} \tag{54}$$

Then we have following properties. For $l\in[L-1]$, $\|h_l-\check{h}_l\|\le O(\tau^2L\omega)$. This is because $\|(\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{a+1})(\check{\boldsymbol{D}}_a+\boldsymbol{D}''_a)\|\le 1.1$ from Lemma 3; $\|h_{a-1}\|\le O(1)$ from Theorem 2; and the assumption $\|\boldsymbol{W}'_l\|\le\tau\omega$ for $l\in[L-1]$.

To have a tighter bound on $\|h_L-\check{h}_L\|$, let us introduce $\boldsymbol{W}''_b:=\sum_{a=b}^{l}(\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{a+1})(\check{\boldsymbol{D}}_a+\boldsymbol{D}''_a)\boldsymbol{W}'_a$, for $b=1,...,l$. Then we have

$$h_L-\check{h}_L=\big[\boldsymbol{W}''_L,\boldsymbol{W}''_{L-1},...,\boldsymbol{W}''_1\big][h_{L-1}^T,\tau h_{L-2}^T,...,\tau h_0^T]^T. \tag{55}$$

It is easy to get

$$\|[\tau h_{l-1}^T,\tau h_{l-2}^T,...,\tau h_0^T]^T\|=\sqrt{\tau^2\sum_{a=0}^{l-1}\|h_a\|^2}\le\tau\sqrt{L}\cdot O(1),$$

where the inequality is because of $\|h_{a-1}\|\le O(1)$ from Theorem 2. Next, we have

$$\Big\|\big[\boldsymbol{W}''_l,\boldsymbol{W}''_{l-1},...,\boldsymbol{W}''_1\big]\Big\|=\Big\|\big[\boldsymbol{W}''_l,\boldsymbol{W}''_{l-1},...,\boldsymbol{W}''_1\big]^T\Big\|\le\sqrt{\sum_{a=1}^{l}\|(\boldsymbol{W}''_l)^T\|^2}\le 1.1\sqrt{\sum_{a=1}^{l}\|(\boldsymbol{W}'_l)^T\|^2}\le 1.1\|\boldsymbol{W}'_{l:1}\|_F, \tag{56}$$

where the second inequality is from the definition of spectral norm, the third inequality is because of $\|(\check{\boldsymbol{D}}_l+\boldsymbol{D}''_l)(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_l)\cdots(\boldsymbol{I}+\tau\check{\boldsymbol{W}}_{a+1})(\check{\boldsymbol{D}}_a+\boldsymbol{D}''_a)\|\le 1.1$ from Lemma 3.

Hence we have $\|h_L-\check{h}_L\|\le O\Big((1+\tau\sqrt{L})\|\boldsymbol{W}'\|_F\Big)=O\big(\|\boldsymbol{W}'\|_F\big)$ because of the assumption $\tau\sqrt{L}\le 1$.

For $l\in[L]$, $\|\boldsymbol{D}''_l\|_0\le O(m\omega^{\frac{2}{3}})$. This is because $(\boldsymbol{D}''_l)_{k,k}$ is non-zero only at coordinates k where $(\check{g}_l)_k$ and $(g_l)_k$ have opposite signs, where it holds either $(\boldsymbol{D}_l^{(0)})_{k,k}\neq(\check{\boldsymbol{D}}_l)_{k,k}$ or $(\boldsymbol{D}_l^{(0)})_{k,k}\neq(\boldsymbol{D}_l)_{k,k}$. Therefore by Lemma 4, we have $\|\boldsymbol{D}''_l\|_0\le O(m(\omega\tau L)^{\frac{2}{3}})$ if $\|\boldsymbol{W}'_l\|\le\tau\omega$.

□

F Proof for Theorem 5

F.1 Convergence result for GD

Proof Using Theorem 2 we have $\|h_{i,L}^{(0)}\| \leq 1.1$ and then using the randomness of $\boldsymbol{B}$, it is easy to show that $\|\boldsymbol{B}h_{i,L}^{(0)} - y_i^*\|^2 \leq O(\log^2 m)$ with probability at least $1 - \exp(-\Omega(\log^2 m))$, and therefore

$$F(\overrightarrow{\boldsymbol{W}}^{(0)}) \leq O(n \log^2 m). \tag{57}$$

Assume that for every $t = 0, 1, \ldots, T-1$, the following holds,

$$\|\boldsymbol{W}_L^{(t)} - \boldsymbol{W}_L^{(0)}\|_F \leq \omega \overset{\Delta}{=} O\left(\frac{\delta^{3/2}}{n^3 L^{7/2}}\right) \tag{58}$$

$$\|\boldsymbol{W}_l^{(t)} - \boldsymbol{W}_l^{(0)}\|_F \leq \tau\omega. \tag{59}$$

We shall prove the convergence of GD under the assumption (58) holds, so that previous statements can be applied. At the end, we shall verify that (58) is indeed satisfied.

Letting $\nabla_t = \nabla F(\overrightarrow{\boldsymbol{W}}^{(t)})$, we calculate that

$$\begin{aligned} F(\overrightarrow{\boldsymbol{W}}^{(t+1)}) &\leq F(\overrightarrow{\boldsymbol{W}}^{(t)}) - \eta\|\nabla_t\|_F^2 + O(\eta^2 nm/d)\|\nabla_t\|_F^2 + \quad \eta\sqrt{F(\overrightarrow{\boldsymbol{W}}^{(t)})} \cdot O\left(\sqrt{\frac{mnL\omega^{\frac{2}{3}}}{d}}(\tau L)^{\frac{4}{3}}\right) \cdot \|\nabla_t\|_F \\ &\leq \left(1 - \Omega\left(\frac{\eta\delta m}{dn}\right)\right) F(\overrightarrow{\boldsymbol{W}}^{(t)}), \end{aligned} \tag{60}$$

where the first inequality uses Theorem 4, the second inequality uses the gradient upper bound in Theorem 6 and the last inequality uses the gradient lower bound in Theorem 7 and the choice of $\eta = O(d/(mn))$ and the assumption on ω (58). That is, after $T = \Omega(\frac{dn}{\eta\delta m}) \log \frac{n \log^2 m}{\epsilon}$ iterations $F(\overrightarrow{\boldsymbol{W}}^{(T)}) \leq \epsilon$.

We need to verify for each t, (58) holds. Here we use a result from the Lemma 4.2 in Zou and Gu (2019) that states $\|\boldsymbol{W}_L^{(t)} - \boldsymbol{W}_L^{(0)}\|_F \leq O(\sqrt{\frac{n^2 d \log m}{m\delta}})$.

To guarantee the iterates fall into the region given by ω (58), we obtain a bound $m \geq n^8 \delta^{-4} dL^7 \log^2 m$. □

F.2 Convergence result for SGD

Theorem 9 *For the ResNet defined and initialized as in Sect. 2, the network width* $m \geq \Omega(n^{17} L^7 b^{-4} \delta^{-8} d \log^2 m)$. *Suppose we do stochastic gradient descent update starting from* $\overrightarrow{\boldsymbol{W}}^{(0)}$ *and*

$$\overrightarrow{\boldsymbol{W}}^{(t+1)} = \overrightarrow{\boldsymbol{W}}^{(t)} - \eta \frac{n}{|S_t|} \sum_{i \in S_t} \nabla F_i(\overrightarrow{\boldsymbol{W}}^{(t)}), \tag{61}$$

where S_t is a random subset of $[n]$ with $|S_t| = b$. Then with probability at least $1 - \exp(-\Omega(\log^2 m))$, stochastic gradient descent (61) with learning rate $\eta = \Theta(\frac{db\delta}{n^3 m \log m})$ finds a point $F(\overrightarrow{\boldsymbol{W}}) \le \epsilon$ in $T = \Omega(n^5 b^{-1}\delta^{-2}\log m \log^2 \frac{1}{\epsilon})$ iterations.

Proof The proof of the case of SGD can be adapted from the proof of Theorem 3.8 in Zou and Gu (2019). □

G Proofs of Theorem 4 and Proposition 1

Proof By induction we can show for any $k \in [m]$ and $l \in [L-1]$,

$$(h_l)_k \ge \phi\left(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right). \tag{62}$$

It is easy to verify $(h_1)_k = \phi\big((h_0)_k + (\tau \boldsymbol{W}_1 h_0)_k\big) \ge \phi\big((\tau \boldsymbol{W}_1 h_0)_k\big)$ because of $(h_0)_k \ge 0$.

Then assume $(h_l)_k \ge \phi\Big(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\Big)$, we show it holds for $l+1$.

$$(h_{l+1})_k = \phi\big((h_l)_k + (\tau \boldsymbol{W}_{l+1} h_l)_k\big) \ge \phi\left(\phi\left(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right) + (\tau \boldsymbol{W}_{l+1} h_l)_k\right) \ge \phi\left(\sum_{a=1}^{l+1} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right),$$

where the last inequality can be shown by case study.

Next we can compute the mean and variance of $\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k$ by taking iterative conditioning. We have

$$\mathbb{E}\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k = 0, \quad \mathbb{E}\left(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right)^2 = \frac{\tau^2}{m} \sum_{a=1}^{l} \mathbb{E}\|h_{a-1}\|^2. \tag{63}$$

Moreover, $(\tau \boldsymbol{W}_a h_{a-1})_k$ are jointly Gaussian for all a with mean 0 because $\boldsymbol{W}_a$'s are drawn from independent Gaussian distributions. We use $l = 2$ as an example to illustrate the conclusion, it can be generalized to other l. Assume that h_0 is fixed. First it is easy to verify that $(\tau \boldsymbol{W}_1 h_0)_k$ is Gaussian variable with mean 0 and $(\tau \boldsymbol{W}_2 h_1)_k \big| \boldsymbol{W}_1$ is also Gaussian variable with mean 0. Hence $[(\tau \boldsymbol{W}_1 h_0)_k, (\tau \boldsymbol{W}_2 h_1)_k]$ follows jointly Gaussian with mean vector [0, 0]. Thus $(\tau \boldsymbol{W}_1 h_0)_k + (\tau \boldsymbol{W}_2 h_1)_k$ is Gaussian with mean 0. By induction, we have $\sum_{a=1}^{l} (\tau \boldsymbol{W}_a h_{a-1})_k$ is Gaussian with mean 0. Then we have

$$\begin{aligned} \mathbb{E}\|h_l\|^2 &\ge \sum_{k=1}^{m} \mathbb{E}\left(\phi\left(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right)\right)^2 = \sum_{k=1}^{m} \frac{1}{2}\mathbb{E}\left(\sum_{a=1}^{l} \left(\tau \boldsymbol{W}_a h_{a-1}\right)_k\right)^2 \\ &= \frac{1}{2}\sum_{k=1}^{m} \frac{\tau^2 \sum_{a=1}^{l} \mathbb{E}\left[\|h_{a-1}\|^2\right]}{m} = \frac{\tau^2}{2}\sum_{a=1}^{l} \mathbb{E}\|h_{a-1}\|^2, \end{aligned} \tag{64}$$

where the first step is due to (62), the second step is due to the symmetry of Gaussian distribution and the third step is due to (66). Since $(h_l)_k = \phi\big((h_{l-1})_k + \left(\boldsymbol{W}_l h_{l-1}\right)_k\big)$, we can show $\mathbb{E}(h_l)_k^2 \ge (h_{l-1})_k^2$ given h_{l-1} by numerical integral of Gaussian variable over an interval. Hence we have $\mathbb{E}\|h_l\|^2 \ge \mathbb{E}\|h_{l-1}\|^2 \ge \cdots \ge \mathbb{E}\|h_0\|^2 = 1$ by iteratively taking conditional expectation. Then combined with (64) and the choice of $\tau = L^{-\frac{1}{2}+c}$, we

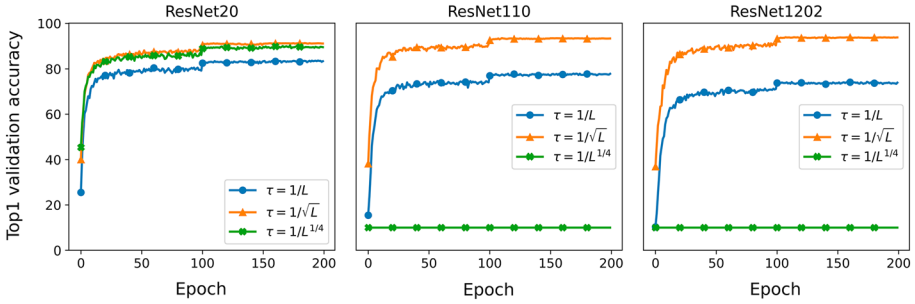

Fig. 6 Validation accuracy on CIFAR10 of ResNets with different choices of τ ($\tau = 1/L$, $\tau = 1/\sqrt{L}$, $\tau = 1/L^{1/4}$)

Table 3 Validation accuracy of ResNet110+τ with different learning rates

Lr	$\tau = 1/L$	$\tau = 1/\sqrt{L}$
0.1	82.7	92.2
0.2	85.6	**92.5**
0.4	**86.8**	92.2
0.8	86.3	90.7
1.6	84.4	10.0

The best accuracy of each column is in bold

have $\mathbb{E}\|h_{L-1}\|^2 \geq \frac{1}{2}L^{2c}$. Because $(\boldsymbol{W}_L)_{i,j} \sim \mathcal{N}(0, 2/m)$ and $h_L = \phi(\boldsymbol{W}_L h_{L-1})$, we have $\mathbb{E}\|h_L\|^2 = \|h_{L-1}\|^2$. Thus, the claim is proved. □

Proof From the inequality (62) in the previous proof, we know for any $k \in [m]$ and $l \in [L-1]$,

$$(h_l)_k \geq \phi\left(\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k\right). \tag{65}$$

Next we can compute the mean and variance of $\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k$ by taking iterative conditioning. We have

$$\mathbb{E}\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k = 0, \quad \mathbb{E}\left(\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k\right)^2 = \sum_{a=1}^{l} \mathbb{E}((\tilde{z}_a)_k)^2 = l. \tag{66}$$

Then we have

$$\mathbb{E}\|h_l\|^2 \geq \sum_{k=1}^{m} \mathbb{E}\left(\phi\left(\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k\right)\right)^2 = \frac{1}{2}\sum_{k=1}^{m} \mathbb{E}\left[\sum_{a=1}^{l} \left(\tilde{z}_a\right)_k\right]^2 = \frac{1}{2}ml, \tag{67}$$

where the first step is due to (62), the second step is due to the symmetry of random variable $(\tilde{z}_a)_k$ and the third step is due to (66). The proposition is proved. □

H More empirical studies

We do more experiments to demonstrate the points in Sect. 5.

Besides the basic feedforward structure in Sect. 5.1, we do another experiment to demonstrate that $\tau = 1/\sqrt{L}$ is sharp with practical structures (see Fig. 6). We can see that for ResNet110 and ResNet1202, $\tau = 1/L^{1/4}$ cannot train the network effectively.

One may wonder if we can tune the learning rate for the case of $\tau = 1/L$ to achieve validation accuracy as well as the case of $\tau = 1/\sqrt{L}$. We do a new experiment to verify this (see Table 3). Specifically, for ResNet110 with fixed $\tau = 1/L$ and $\tau = 1/\sqrt{L}$ on CIFAR10 classification task, we tune the learning rate from 0.1 to 1.6 and record the validation accuracy in Table 3. We can see that ResNet110 with $\tau = 1/L$ performs inferior to that with $\tau = 1/\sqrt{L}$ even with grid search of learning rates. It is possible that we can achieve a bit better performance by adjusting the learning rate for $\tau = 1/L$. But this requires tuning for each depth. In contrast, we have shown that with $\tau = 1/\sqrt{L}$, one learning rate fits for all nets with different depths.

References

Allen-Zhu, Z., & Li, Y. (2019). What can ResNet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*.

Allen-Zhu, Z., Li, Y., & Song, Z. (2018). A convergence theory for deep learning via over-parameterization. arXiv preprint arXiv:1811.03962.

Allen-Zhu, Z., Li, Y., & Liang, Y. (2019a). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in Neural Information Processing Systems*, pp.6155–6166.

Allen-Zhu, Z., Li, Y., & Song, Z. (2019b). On the convergence rate of training recurrent neural networks. *Advances in Neural Information Processing Systems*.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., & Wang, R. (2019a). On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*.

Arora, S., Du, S. S., Hu, W., Li, Z., & Wang, R. (2019b). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *International Conference on Machine Learning (ICML)*.

Arpit, D., Campos, V., & Bengio, Y. (2019). How to initialize your network? robust initialization for weightnorm & resnets. *Advances in Neural Information Processing Systems*.

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Wan-Duo Ma, K., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning (ICML)*, pp. 342–350.

Brutzkus, A., Globerson, A., Malach, E., & Shalev-Shwartz, S. (2018). SGD learns over-parameterized networks that provably generalize on linearly separable data. In *Proceedings of the 6th international conference on learning representations (ICLR 2018)*.

Cao, Y., & Gu, Q. (2019). A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. arXiv preprint arXiv:1902.01384.

Cao, Y., & Gu, Q. (2020). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, Z., Cao, Y., Zou, D., & Gu, Q. (2021). How much over-parameterization is sufficient to learn deep ReLU networks? In *Proceedings of the international conference on learning representations (ICLR 2021)*.

Chizat, L., & Bach, F. (2018a). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems 31*.

Chizat, L., & Bach, F. (2018b). A note on lazy training in supervised differentiable programming. arXiv preprint arXiv:1812.07956, 8.

Chizat, L., Oyallon, E., & Bach, F. (2019). On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2019a). Gradient descent finds global minima of deep neural networks. In: *International Conference on Machine Learning (ICML)*.
Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2019b). Gradient descent provably optimizes over-parameterized neural networks. In: *International Conference on Learning Representations (ICLR)*.
Fang, C., Dong, H., & Zhang, T. (2019a). Over parameterized two-level neural networks can learn near optimal feature representations. arXiv preprint arXiv:1910.11508.
Fang, C., Gu, Y., Zhang, W., & Zhang, T. (2019b). Convex formulation of overparameterized deep neural networks. arXiv preprint arXiv:1911.07626.
Frei, S., Cao, Y., & Gu, Q. (2019). Algorithm-dependent generalization bounds for overparameterized deep residual networks. *Advances in Neural Information Processing Systems*, pages 14769–14779.
Ghorbani, B., Mei, S., Misiakiewicz, T., Montanari, A. (2019). Limitations of lazy training of two-layers neural networks. *Advances in Neural Information Processing Systems*.
Haber, E., & Ruthotto, L. (2017). Stable architectures for deep neural networks. *Inverse Problems, 34*(1), 014004.
Hardt, M., & Ma, T. (2016). Identity matters in deep learning. In: *International Conference on Learning Representations (ICLR)*.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning (ICML)*, pp. 448–456.
Jacot, A., Gabriel, F, & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, pp. 8571–8580.
Ji, Z., & Telgarsky, M. (2020). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *Proceedings of the international conference on learning representations (ICLR 2020)*.
Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338.
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, pp. 8168–8177.
Mei, S., Montanari, A., & Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences, 115*(33), E7665–E7671.
Mei, S., Misiakiewicz, T., & Montanari, A. (2019). Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the thirty-second conference on learning theory* (pp. 2388–2464).
Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., & Srebro, N. (2019). The role of over-parametrization in generalization of neural networks. In: *International Conference on Learning Representations (ICLR)*.
Nguyen. P.-M. (2019). Mean field limit of the learning dynamics of multilayer neural networks. arXiv preprint arXiv:1902.02880.
Orhan, A. E., & Pitkow, X. (2018). Skip connections eliminate singularities. In: *International Conference on Learning Representations (ICLR)*.
Oymak, S., & Soltanolkotabi, M. (2019). Overparameterized nonlinear learning: Gradient descent takes the shortest path? In: *International Conference on Machine Learning (ICML)*.
Spielman, D. A., & Teng, S-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM), 51*(3):385–463.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. *Advances in Neural Information Processing Systems*, pp 550–558.
Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* (pp. 210–268). Theory and Applications: Compressed Sensing.
Yang, G., and Schoenholz, S. (2017). Mean field residual networks: On the edge of chaos. *Advances in Neural Information Processing Systems*, pp 7103–7114.
Yehudai, G., & Shamir, O. (2019). On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*.

Zhang, H., Dauphin, Y. N., & Ma, T. (2019a). Fixup initialization: Residual learning without normalization. In: *International Conference on Learning Representations (ICLR)*.
Zhang, H., Chen, W., & Liu, T.-Y. (2018). On the local hessian in back-propagation. In *Advances in Neural Information Processing Systems*, pp. 6521–6531.
Zhang, J., Han, B., Wynter, L., Low, K. H., & Kankanhalli, M. (2019b). Towards robust resnet: A small step but a giant leap. In: *International Joint Conferences on Artificial Intelligence (IJCAI)*.
Zou, D., & Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. *Advances in Neural Information Processing Systems*.
Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2020). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, *109*(3), 467–492.