# Supplementary Materials for: Gradient Descent Optimizes Normalization-Free ResNets

Zongpeng Zhang, Zenan Ling, Tong Lin, Zhouchen Lin

## I. Theoretical Results

**Proposition 1:** Assume that $\sigma(\cdot)$ satisfies Assumption 2, and for any $i, j \in [n]$, $x_i$ and $x_j$ are not parallel. Then we have

$$\lambda_0 > 0.$$

**Lemma 1 (The full rankness of $\mathbf{G}^{(H)}(0)$):** Assume that the number $m$ of neurons per layer is $\Omega\left(\frac{n^2\log(Hn/\delta)}{\lambda_0^2}\right)$, then with probablity at least $1-\delta$ we have:

$$\lambda_{\min}(\mathbf{G}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.$$

**Lemma 2 (The full rankness of $\mathbf{G}^{(H)}(k)$):** Assume that the number $m$ of neurons per layer is $\Omega\left(\frac{n^2\log(Hn/\delta)}{\lambda_0^2}\right)$, then with probablity at least $1-\delta$ we have:

$$\lambda_{\min}(\mathbf{G}^{(H)}(k)) \geq \frac{\lambda_0}{4}.$$

**Theorem 1 (The convergence of the loss):** Assume for all $i \in [n]$, $\|x_i\|_2 = 1$, $y_i = O(1)$, $m = \Omega\Bigg(\max\Bigg\{\frac{n^4}{\lambda_0^4 H^6}, \frac{n^2}{\lambda_0^2 H^2}, \frac{n}{\delta}, \frac{n^2\log(Hn/\delta)}{\lambda_0^2}\Bigg\}\Bigg)$ and we set the step size $\eta = O\left(\frac{\lambda_0 H^2}{n^2}\right)$, then with probability at least $1-\delta$ over the random initialization, we have

$$\|y - u(k)\|_2^2 \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k \|y - u(0)\|_2^2, \quad k = 1, 2, \cdots. \tag{1}$$

**Lemma 3 (Bounded Weight Pertubation):** If Assumptions 1 and 2 hold and $\eta \leq c\frac{H^2}{m}$ for some small constant $c > 0$, then we have

$$\|\alpha_h(k) - \alpha_h(0)\|_F \leq O(1),$$
$$\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq O(\sqrt{n}), \quad k = 0, 1, 2, \cdots.$$

**Lemma 4 (Bounded Output Pertubation):** Suppose that $\sigma(\cdot)$ is $L$-Lipschitz and for $h \in [H]$, $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$, $\left\|x^{(h)}(0)\right\|_2 \leq c_{x,0}$, $\|a_h(0)\|_2 \leq c_{a,0}$ and $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0}, c_{a,0} > 0$ and $R \leq c_{w,0}$. Then with probability at least $1-\delta$, we have

$$\left\|x^{(h)}(k) - x^{(h)}(0)\right\|_2 = O(1).$$

**Lemma 5 (The boundedness of initial output):** If $m = \Omega\left(\frac{n}{\delta}\right)$, we have with probability at least $1-\delta$ over random initialization that

$$\frac{1}{c_{x,0}} \leq \|x_i^{(h)}(0)\|_2 \leq c_{x,0}, \text{ for all } h \in [H] \text{ and } i \in [n].$$

for some universal constant $c_{x,0} > 1$ (only depending on $\sigma$).

**Lemma 6:** If Theorem 1 holds for $k' = 1, \cdots, k$, we have

$$\|I_2(k)\|_2 = O(\eta^2 \|y - u(k)\|_2).$$

First, we give an intermediate conclusion.

**Lemma 7:** $(y-u(k))^\top I_1(k) \geq \lambda_{\min}(\mathbf{G}^{(H)}(k))\|y-u(k)\|_2^2$.

**Lemma 8:** If Theorem 1 holds for $k' = 1, \cdots, k$ and $\eta \leq c\lambda_0 H^2 n^{-2}$ for some small constant $c$, then

$$\|u(k+1) - u(k)\|_2^2 = O(\eta^2 \|y - u(k)\|_2^2).$$

**Theorem 2 (No vanishing or exploding gradient throughout training):** There exists a lower bound $m_1$ and upper bound $M$ of the magnitude of gradients during training for all iteration $k$, such that

$$m_1 \leq \left\|\frac{\partial u_i}{\partial \xi}\right\|_2 \leq M, \quad \text{for all } i \in [n], k \in \mathbb{N} \text{ and } \xi \in \theta,$$

where $m_1 > 0$ and $M = O(\sqrt{n})$.

**Theorem 3 (Exploding gradient of standard initialization):** For ResNets with standard initialization, the condition number of $\mathbf{J}\mathbf{J}^\top$ grows at least linearly with depth. More specifically,

$$\lambda_{\max}(\mathbf{J}\mathbf{J}^\top) = \Omega(H),$$
$$\lambda_{\min}(\mathbf{J}\mathbf{J}^\top) = O(1).$$

## II. Proof of Main Results

The gradient for the normalization-free ResNet structure (1) is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(h)}} = \frac{\alpha_h}{H\sqrt{m}} \sum_{i=1}^{n} (y_i - u_i) x_i^{(h-1)} \cdot \left[ a^\top \prod_{l=h+1}^{H} \left( I + \frac{\alpha_l}{H\sqrt{m}} \mathbf{J}_i^{(l)} \mathbf{W}^{(l)} \right) \mathbf{J}_i^{(h)} \right],$$

where

$$\mathbf{J}^{(h')} \triangleq diag\left( \sigma'\left( (w_1^{(h')})^\top x^{(h'-1)} \right), \cdots, \sigma'\left( (w_m^{(h')})^\top x^{(h'-1)} \right) \right) \in \mathbb{R}^{m \times m}.$$

For our network, the entries of $\mathbf{G}^{(H)}$ have the following form:

$$\mathbf{G}_{ij}^{(H)} = \frac{(\alpha_H)^2}{H^2 m} (x_i^{(H-1)})^\top x_j^{(H-1)} \sum_{r=1}^{m} a_r^2 \sigma'\left( (w_r^{(H)})^\top x_i^{(H-1)} \right) \sigma'\left( (w_r^{(H)})^\top x_j^{(H-1)} \right). \tag{2}$$

Then we describe our main idea of proving the global convergence of the structure (1).

We use mathematical induction to prove the conclusion. Our induction hypothesis is just the convergence rate in Theorem 1 of empirical loss. At the $(k+1)$-th step, the loss is given by

$$\|y - u(k+1)\|_2^2 = \|y - u(k) - (u(k+1) - u(k))\|_2^2$$
$$= \|y - u(k)\|_2^2 - 2(y - u(k))^\top (u(k+1) - u(k)) + \|u(k+1) - u(k)\|_2^2.$$

We look at one entry of $u(k+1)-u(k)$. Using Taylor expansion, we have

$$u_i(k+1)-u_i(k)=u_i(\theta(k)-\eta\mathcal{L}'(\theta(k)))-u_i(\theta(k))=-\int_0^\eta\langle\mathcal{L}'(\theta(k)),u_i'(\theta(k)-s\mathcal{L}'(\theta(k))\rangle ds$$
$$=-\int_0^\eta\langle\mathcal{L}'(\theta(k)),u_i'(\theta(k))\rangle ds+\int_0^\eta\langle\mathcal{L}'(\theta(k)),u_i'(\theta(k))-u_i'(\theta(k)-s\mathcal{L}'(\theta(k)))\rangle ds\triangleq I_1^i(k)+I_2^i(k).$$

Denote $I_1(k)=(I_1^1(k),\cdots,I_1^n(k))^\top$ and $I_2(k)=(I_2^1(k),\cdots,I_2^n(k))^\top$. Then $u(k+1)-u(k)=I_1(k)+I_2(k)$. Then we can prove that $(y-u(k))^\top I_1(k)\geq\lambda_{\min}(\mathbf{G}^{(H)}(k))\|y-u(k)\|_2^2$ (Lemma 7), so there is

$$\begin{aligned}&\|y-u(k+1)\|_2^2\\ \leq&(1-2\eta\lambda_{\min}(\mathbf{G}^{(H)}(k))\|y-u(k)\|_2^2-2(y-u(k))^\top I_2(k)+\|u(k+1)-u(k)\|_2^2.\end{aligned}\tag{3}$$

Combining the above formula and Lemmas 2, 6 and 8, we can get the main theorem. These results bound three terms of Equation (3).

Now we proceed to analyze the training process.

We prove the following lemma which characterizes how the perturbation from weight matrices propagates to the input of each layer. This lemma is used to prove the subsequent lemmas.

*Proof of Lemma 2:*

Firstly, we can prove that if $m=\Omega\left(\frac{n^2\log(Hn/\delta)}{\lambda_0^2}\right)$,

$$\left\|\mathbf{G}^{(H)}(0)-\mathbf{K}^{(H)}\right\|\leq\frac{\lambda_0}{4},$$

which is a direct consequence of results in Section E of [1].

Secondly, suppose that $\sigma(\cdot)$ fulfils Assumption 2. Suppose for $h\in[H]$,

$$\|\mathbf{W}^{(h)}(0)\|_2\leq c_{w,0}\sqrt{m},\|a(0)\|_2\leq a_{2,0}\sqrt{m},\|a(0)\|_4\leq a_{4,0}m^{\frac{1}{4}},$$
$$\frac{1}{c_{x,0}}\leq\|x^{(h)}(0)\|\leq c_{x,0},\quad\text{if}\quad\|\mathbf{W}^{(h)}(k)-\mathbf{W}^{(h)}(0)\|_F,\|\alpha_h(k)-\alpha_h(0)\|_2\leq mR,$$

where $R\leq c\lambda_0H^2n^{-1}$ and $R\leq c$ for some small constant $c$, we have

$$\left\|\mathbf{G}^{(H)}(k)-\mathbf{G}^{(H)}(0)\right\|_2\leq\frac{\lambda_0}{2}.$$

From these two points, we can get $\left\|\mathbf{G}^{(H)}(k)-\mathbf{K}^{(H)}(0)\right\|\leq\frac{3}{4}\lambda_0$. Then using the Hoffman-Wielandt theorem [2], the lemma is proven.

So the key is to prove the second point.

Because the Frobenius-norm of a matrix is greater than the operator norm, it is sufficient to bound $\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\|_F$. For simplicity, define $z_{i,r}(k) = w_r^{(H)}(k)^\top x_i^{(H-1)}(k)$, then we have

$$\begin{aligned}
&\left|\mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0)\right| \\
=&\left|\frac{\alpha_h^2(k)}{H^2 m}(x_i^{(H-1)}(k))^\top x_j^{(H-1)}(k)\sum_{r=1}^{m} a_r^2(k)\sigma'(z_{i,r}(k))\sigma'(z_{j,r}(k))\right. \\
&\left.-\frac{\alpha_h^2(0)}{H^2 m}(x_i^{(H-1)}(0))^\top x_j^{(H-1)}(0)\sum_{r=1}^{m} a_r^2(0)\sigma'(z_{i,r}(0))\sigma'(z_{j,r}(0))\right| \\
=&\left|(x_i^{(H-1)}(k))^\top x_j^{(H-1)}(k)\sum_{r=1}^{m}\frac{\alpha_h^2(k)}{H^2 m} a_r^2(k)\sigma'(z_{i,r}(k))\sigma'(z_{j,r}(k))\right. \\
&\left.-x_i^{(H-1)}(0))^\top x_j^{(H-1)}(0)\sum_{r=1}^{m}\frac{\alpha_h^2(0)}{H^2 m} a_r^2(0)\sigma'(z_{i,r}(0))\sigma'(z_{j,r}(0))\right| \\
\leq&\frac{\alpha_{\max}^2}{H^2}\left(I_1^{i,j} + I_2^{i,j} + I_3^{i,j}\right).
\end{aligned}$$

For $I_1^{i,j}$, using Lemma 4, we have

$$\begin{aligned}
I_1^{i,j} =& L^2 a_{2,0}^2 \left|x_i^{(H-1)}(k)^\top x_j^{(H-1)}(k) - x_i^{(H-1)}(0)^\top x_j^{(H-1)}(0)\right| \\
\leq& L^2 a_{2,0}^2 \left|(x_i^{(H-1)}(k) - x_i^{(H-1)}(0))^\top x_j^{(H-1)}(k)\right| \\
&+ L^2 a_{2,0}^2 \left|x_i^{(H-1)}(0)^\top (x_i^{(H-1)}(k) - x_i^{(H-1)}(0))\right| \\
\leq& c_x L^2 a_{2,0}^2 R \cdot (c_{x,0} + c_x R) + c_{x,0} c_x L^2 a_{2,0}^2 R \\
\leq& 3 c_{x,0} c_x L^2 a_{2,0}^2 R,
\end{aligned}$$

where $c_x \triangleq \left(\sqrt{c_\sigma} L + \frac{c_{x,0}}{c_{w,0}}\right) e^{2\alpha_{\max} c_{w,0} L}$. To bound $I_2^{i,j}$, we have

$$\begin{aligned}
I_2^{i,j} =& c_{x,0}^2 \frac{1}{m}\left|\sum_{r=1}^{m} a_r(0)^2 \sigma'\left(z_{i,r}(k)\right)\sigma'\left(z_{j,r}(k)\right) - a_r(0)^2 \sigma'\left(z_{i,r}(0)\right)\sigma'\left(z_{j,r}(0)\right)\right| \\
\leq& c_{x,0}^2 \frac{1}{m}\sum_{r=1}^{m} a_r(0)^2 \left|\left(\sigma'\left(z_{i,r}(k)\right) - \sigma'\left(z_{i,r}(0)\right)\right)\sigma'\left(z_{j,r}(k)\right)\right| \\
&+ a_r(0)^2 \left|\left(\sigma'\left(z_{j,r}(k)\right) - \sigma'\left(z_{j,r}(0)\right)\right)\sigma'\left(z_{i,r}(0)\right)\right| \\
\leq& \frac{\beta L c_{x,0}^2}{m}\left(\sum_{r=1}^{m} a_r(0)^2 |z_{i,r}(k) - z_{i,r}(0)| + a_r(0)^2 |_{j,r}(k) - z_{j,r}(0)|\right) \\
\leq& \frac{\beta L a_{4,0}^2 c_{x,0}^2}{\sqrt{m}}\left(\sqrt{\sum_{r=1}^{m} |z_{i,r}(k) - z_{i,r}(0)|^2} + \sqrt{\sum_{r=1}^{m} |z_{j,r}(k) - z_{j,r}(0)|^2}\right).
\end{aligned}$$

Using the same proof for Lemma 4, it is easy to see that

$$\sum_{r=1}^{m} |z_{i,r}(k) - z_{i,r}(0)|^2 \leq \left(2 c_x c_{w,0} + c_{x,0}\right)^2 L^2 m R^2.$$

Thus

$$I_2^{i,j} \leq 2\beta c_{x,0}^2 \left(2 c_x c_{w,0} + c_{x,0}\right) L^2 R.$$

Similarly, we can prove,

$$I_3^{i,j} \leq 12L^2 c_{x,0}^2 a_{2,0} R.$$

Therefore, we can bound the perturbation

$$\begin{aligned}\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_F = & \sqrt{\sum_{(i,j)}^{n,n} \left|\mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0)\right|^2} \\ \leq & \frac{\alpha_{\max}^2 L^2 nR}{H^2}\left[3c_{x,0}c_x a_{2,0}^2 + 2\beta c_{x,0}^2\left(2c_x c_{w,0} + c_{x,0}\right)a_{4,0}^2 + 12c_{x,0}^2 a_{2,0}\right].\end{aligned}$$

Plugging in the bound on $R$, we have the desired result. ■

*Proof of Lemma 3:* We will prove this lemma by induction. The induction hypothesis is

$$\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sum_{s'=0}^{s-1}\left(1 - \frac{\eta\lambda_0}{2}\right)^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{n} \leq R'\sqrt{n}, s \in [k+1],$$

$$\|\alpha_h(s) - \alpha_h(0)\|_2 \leq \sum_{s'=0}^{s-1}\left(1 - \frac{\eta\lambda_0}{2}\right)^{s'/2}\frac{1}{4}\eta\lambda_0 M' \leq M', s \in [k+1],$$

where $R' = \frac{16\alpha_h c_{x,0} a_{2,0} L e^{2\alpha_h c_{w,0} L}\sqrt{n}\|y-u(0)\|_2}{\lambda_0 H\sqrt{m}}$, $M' = \frac{32 a_{2,0} C_{x,0} C_{w,0} e^{2C_{w,0}}\|y-u(0)\|_2\sqrt{n}}{\lambda_0}$. As $m$ is much larger than $n$ in over-parameterized ResNets, $\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F$ can be upper bounded by $C\sqrt{n}$ for an absolute constant $C$.

First it is easy to see it holds for $s' = 0$. Now suppose it holds for $s' = 0, \cdots, s$, we consider $s' = s+1$. We have

$$\begin{aligned}&\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F \\ \leq & \eta\frac{\alpha_h}{H\sqrt{m}}\|a\|_2\sum_{i=1}^n |y_i - u_i(s)|\left\|x_i^{(h-1)}(s)\right\|_2\prod_{k=h+1}^H\left\|I + \frac{\alpha_h}{H\sqrt{m}}\mathbf{J}_i^{(k)}(s)\mathbf{W}^{(k)}(s)\right\|_2 \\ \leq & 2\eta\alpha_h c_{x,0} L a_{2,0} e^{2c_{w,0}}\frac{n}{H\sqrt{m}}\|y - u(s)\|_2 \\ = & \eta Q^{'}(s) \\ \leq & \left(1 - \frac{\eta\lambda_0}{2}\right)^{s/2}\frac{1}{4}\eta\lambda_0 R^{'}\sqrt{n}.\end{aligned}$$

Thus

$$\begin{aligned}&\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(0)\right\|_F \\ \leq & \left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F + \left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F \\ \leq & \sum_{s'=0}^{s}\eta\left(1 - \frac{\eta\lambda_0}{2}\right)^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{n} \\ = & O(\sqrt{n}).\end{aligned}$$

Similarly, we have $\|a(s)-a(0)\|_2 = O(1)$.

Similarly, because $\frac{\partial L}{\partial \alpha_h} = -\sum_{i=1}^{n}(y-u_i)a_i \cdot \prod_{l=h+1}^{H}(I + \frac{\alpha_l}{H\sqrt{(m)}}\mathbf{J}_i^{(l)}\mathbf{W}^{(l)}\frac{1}{H\sqrt{m}})$, when $k=1$ we have

$$
\begin{aligned}
\|\alpha_h(1)-\alpha_h(0)\|_2 &= \left\|\eta\frac{\partial L}{\partial \alpha_h(0)}\right\|_2 \\
\leq &\eta\left\|\sum_{i=1}^{n}(y-u_i)a_i\right\| \cdot \prod_{l=h+1}^{H}\left\|\left(I+\frac{\alpha_l}{H\sqrt{m}}\mathbf{J}_i^{(l)}\mathbf{W}^{(l)}\right)\frac{1}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}x^{(h-1)}\right)\right\|_2 \\
\leq &\eta\|y-u(0)\|_2\|a\|_2 \ \cdot \prod_{l=h+1}^{H}\left\|\left(I+\frac{\alpha_l}{H\sqrt{m}}\mathbf{J}_i^{(l)}\mathbf{W}^{(l)}\right)\frac{1}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}x^{(h-1)}\right)\right\|_2 \\
\leq &\eta\left(1-\frac{\eta\lambda_0}{2}\right)^{0/2}\|y-u(0)\|_2 \cdot 2a_{2,0}e^{2\alpha_h(0)C_{w,0}} \cdot 2C_{w,0} \cdot 2C_{x,0}\frac{\sqrt{n}}{H} \\
= &\left(1-\frac{\eta\lambda_0}{2}\right)^{0/2}\frac{1}{4}\eta\lambda_0 M'\frac{\sqrt{n}}{H} \\
\leq &M'\frac{\sqrt{n}}{H}.
\end{aligned}
$$

More generally,

$$
\begin{aligned}
\|\alpha_h(s+1)-\alpha_h(s)\|_2 &= \left\|\eta\frac{\partial L}{\partial \alpha_h(s)}\right\|_2 \\
\leq &\eta|\sum_{i=1}^{n}(y-u_i)a_i| \cdot \prod_{l=h+1}^{H}\left\|\left(I+\frac{\alpha_l}{H\sqrt{m}}\mathbf{J}_i^{(l)}\mathbf{W}^{(l)}\right)\frac{1}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}x^{(h-1)}\right)\right\|_2 \\
\leq &\eta\|y-u(s)\|_2\|a\|_2 \ \cdot \prod_{l=h+1}^{H}\left\|(I+\frac{\alpha_l}{H\sqrt{m}}\mathbf{J}_i^{(l)}\mathbf{W}^{(l)})\frac{1}{H\sqrt{m}}\sigma(\mathbf{W}^{(h)}x^{(h-1)}))\right\|_2 \\
\leq &\eta\left(1-\frac{\eta\lambda_0}{2}\right)^{s/2}\|y-u(0)\|_2 \cdot 2a_{2,0}e^{2\alpha_h(s)C_{w,0}} \cdot 2C_{w,0} \cdot 2C_{x,0}\frac{\sqrt{n}}{H} \\
= &\left(1-\frac{\eta\lambda_0}{2}\right)^{s/2}\frac{1}{4}\eta\lambda_0 M'\frac{\sqrt{n}}{H}e^{2(\alpha_h(s)-1)C_{w,0}}.
\end{aligned}
$$

We will control $\eta$ small enough, so that the two adjacent $\alpha_h(s)$ does not change much.

Thus, when $k=s$,

$$
\begin{aligned}
\|\alpha_h(s+1)-\alpha_h(0)\| = \sum_{k=1}^{s+1}&|\alpha_h(k)-\alpha_h(k-1)| \\
\leq &\|\alpha_h(s+1)-\alpha_h(s)\|_2 + \|\alpha_h(s)-\alpha_h(0)\|_2 \\
\leq &\left(1-\frac{\eta\lambda_0}{2}\right)^{s/2}\frac{1}{4}\eta\lambda_0 M'\frac{n}{H\sqrt{m}}e^{2(\alpha_h(s)-1)C_{w,0}} \\
&+\sum_{k=0}^{s-1}\left(1-\frac{\eta\lambda_0}{2}\right)^{k/2}\frac{1}{4}\eta\lambda_0 M'\frac{\sqrt{n}}{H} \\
\leq &\sum_{k=0}^{s}\left(1-\frac{\eta\lambda_0}{2}\right)^{k/2}\frac{1}{4}\eta\lambda_0 M'\frac{\sqrt{n}}{H}.
\end{aligned}
$$

The last inequality is because there is $\alpha_h(s) \leq 1 + R'\frac{\sqrt{n}}{H}$, so we can make $e^{\alpha_h}$ bounded.

■

*Proof of Lemma 4:* We prove this lemma by induction. Our induction hypothesis is

$$\|x^{(h)}(k) - x^{(h)}(0)\|_2 \leq g(h),$$

where

$$g(h) = g(h-1)\left(1 + 2\alpha_h c_{w,0} L\right) + \alpha_h(k) R c_{x,0}.$$

For $h = 1$, we have

$$\begin{aligned}\|x^{(1)}(k) - x^{(1)}(0)\|_2 &\leq \sqrt{\frac{c_\sigma}{m}} \|\sigma(\mathbf{W}^{(1)}(k)x) - \sigma(\mathbf{W}^{(1)}(0)x)\|_2 \\ &\leq \sqrt{\frac{c_\sigma}{m}} \|\mathbf{W}^{(1)}(k) - \mathbf{W}^{(1)}(0)\|_F \leq \sqrt{c_\sigma} L R,\end{aligned}$$

which implies $g(1) = \sqrt{c_\sigma} LR$. For $2 \leq h \leq H$, we have

$$\begin{aligned}&\left\|x^{(h)}(k) - x^{(h)}(0)\right\|_2 \\ \leq&\frac{\alpha_h(k)}{H\sqrt{m}} \left\|\sigma(\mathbf{W}^{(h)}(k)x^{(h-1)}(k)) - \sigma(\mathbf{W}^{(h)}(0)x^{(h-1)}(0))\right\|_2 + \left\|x^{(h-1)}(k) - x^{(h-1)}(0)\right\|_2 \\ \leq&\frac{\alpha_h(k)}{H\sqrt{m}} \left\|\sigma(\mathbf{W}^{(h)}(k)x^{(h-1)}(k)) - \sigma(\mathbf{W}^{(h)}(k)x^{(h-1)}(0))\right\|_2 + \frac{\alpha_h(k)}{H\sqrt{m}} \left\|\sigma(\mathbf{W}^{(h)}(k)x^{(h-1)}(0)) - \sigma(\mathbf{W}^{(h)}(0)x^{(h-1)}(0))\right\|_2 \\ &+ \left\|x^{(h-1)}(k) - x^{(h-1)}(0)\right\|_2 \\ \leq&\frac{\alpha_h(k)L}{H\sqrt{m}} \left(\left\|\mathbf{W}^{(h)}(0)\right\|_2 + \left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F\right) \cdot \left\|x^{(h-1)}(k) - x^{(h-1)}(0)\right\|_2 \\ &+ \frac{\alpha_h(k)L}{H\sqrt{m}} \left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \|x^{h-1}(0)\|_2 + \left\|x^{(h-1)}(k) - x^{(h-1)}(0)\right\|_2 \\ \leq& \left[1 + \frac{\alpha_h(k)L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R\sqrt{m})\right] g(h-1) + \frac{\alpha_h(k)}{H} LRc_{x,0} \\ \leq& \left(1 + \frac{2\alpha_h(k)c_{w,0}L}{H}\right) g(h-1) + \frac{\alpha_h(k)}{H} Lc_{x,0}R.\end{aligned}$$

Lastly, simple calculations show that $g(h) \leq (\sqrt{c_\sigma}L + \frac{c_{x,0}}{c_{w,0}})e^{2c_{w,0}}R$. So we get $\left\|x^{(h)}(k) - x^{(h)}(0)\right\|_2 \leq (\sqrt{c_\sigma} + \frac{c_{x,0}}{c_{w,0}})e^{\frac{2\alpha_h c_{w,0}L}{H}} R + e^{\frac{\alpha_h c_{x,0}L}{H}} R$. Substituting the conclusion of Lemma 4 to this conclusion, we have $\left\|x^{(h)}(k) - x^{(h)}(0)\right\|_2 = O(e^{\frac{\alpha_h}{H}}) = O(e^{\frac{1}{H}})$. ■

*Proof of Lemma 5:* First we can prove that given a matrix $\mathbf{W} \in \mathbb{R}^{m \times cm}$ with $\mathbf{W}_{i,j} \sim N(0,1)$, where $c$ is a constant, we have with probability at least $1 - \exp\left(-\frac{\left(c_{w,0} - \sqrt{c} - 1\right)^2 m}{2}\right)$ that

$$\|\mathbf{W}\|_2 \leq c_{w,0}\sqrt{m}, \tag{4}$$

where $c_{w,0} > \sqrt{c} + 1$ is a constant.

The above conclusion is a consequence of well-known deviations bounds concerning the singular values of Gaussian random matrices [3]:

$$P\left(\lambda_{\max}(\mathbf{W}) > \sqrt{m} + \sqrt{cm} + t\right) \leq e^{-t^2/2}.$$

Choosing $t = (c_{w,0} - \sqrt{c} - 1)\sqrt{m}$, we prove (4).

Thus there exists $\|\mathbf{W}^{(h)}(0)\|_2 \leq c_{w,0}\sqrt{m}$ for $h \in [2, H]$ and $c_{w,0} \approx 2$ for Gaussian initialization.

Next, we will bound $\|x_i^{(h)}(0)\|_2$ layer by layer. For the first layer, we can calculate

$$\begin{aligned} E[\|x_i^{(1)}(0)\|_2^2] =& c_\sigma E[\sigma(w_r^{(1)}(0)^\top x_i)^2] \\ =& c_\sigma E_{X\sim N(0,1)}[\sigma(X)^2] \\ =& 1. \end{aligned}$$

$$\begin{aligned} Var[\|x_i^{(1)}(0)\|_2^2] =& \frac{c_\sigma^2}{m} Var[\sigma(w_r^{(1)}(0)^\top x_i(0))^2] \\ \leq& \frac{c_\sigma^2}{m} E_{X\sim N(0,1)}\sigma(X)^4 \\ \leq& \frac{c_\sigma^2}{m} E\left[\left(\left|\sigma(0)\right| + L\left|w_r^{(1)}(0)^\top x_i\right|\right)^4\right] \\ \leq& \frac{C_2}{m}, \end{aligned}$$

where $C_2 \triangleq \sigma(0)^4 + 4|\sigma(0)|^3 L\sqrt{2/\pi} + 6\sigma(0)^2 L^2 + 8|\sigma(0)|L^3\sqrt{2/\pi} + 32L^4$. We have with probability at least $1 - \frac{\delta}{n}$ that,

$$\frac{1}{2} \leq \left\|x_i^{(1)}(0)\right\|_2 \leq 2.$$

At the initial time, we have $\alpha_h = 1$. So by definition, we have for $2 \leq h \leq H$,

$$\begin{aligned} \|x_i^{(h-1)}(0)\|_2 - \left\|\frac{1}{H\sqrt{m}}\sigma(\mathbf{W}^{(h)}(0)x_i^{(h-1)}(0))\right\|_2 &\leq \|x^{(h)}(0)\|_2 \\ &\leq \|x_i^{(h-1)}(0)\|_2 + \left\|\frac{1}{H\sqrt{m}}\sigma(\mathbf{W}^{(h)}(0)x^{(h-1)}(0))\right\|_2, \end{aligned}$$

where $\|\frac{1}{H\sqrt{m}}\sigma(\mathbf{W}^{(h)}(0)x_i^{(h-1)}(0))\|_2 \leq \frac{c_{w,0}L}{H}\|x_i^{(h-1)}(0)\|_2$. Thus

$$\|x_i^{(h-1)}(0)\|_2\left(1 - \frac{c_{w,0}L}{H}\right) \leq \|x_i^{(h)}(0)\|_2 \leq \|x_i^{(h-1)}(0)\|_2\left(1 + \frac{c_{w,0}L}{H}\right),$$

which implies $\frac{1}{2}e^{-c_{w,0}L} \leq \|x^{(h)}(0)\|_2 \leq 2e^{c_{w,0}L}$. Choosing $c_{x,0} = 2e^{c_{w,0}L}$ and using union bounds over $[n]$, we prove the lemma.

■

*Proof of Lemma 6:* We first bound the gradient norm.

$$\begin{aligned} &\left\|\mathcal{L}'^{(h)}(w(k))\right\|_F \\ =& \left\|\frac{\alpha_h}{H\sqrt{m}}\sum_{i=1}^n \left(y_i - u_i(k)\right)x_i^{(h-1)}(k)\cdot\left[a(k)^\top\prod_{l=h+1}^H\left(I + \frac{\alpha_h}{H\sqrt{m}}\mathbf{J}_i^{(l)}(k)\mathbf{W}^{(l)}(k)\right)\mathbf{J}_i^{(h)}(k)\right]\right\|_F \\ \leq& \frac{\alpha_h}{H\sqrt{m}}\|a(k)\|_2\sum_{i=1}^n|y_i - u_i(k)|\cdot\|x^{(h-1)}(k)\|_2\prod_{k=h+1}^H\left\|I + \frac{\alpha_h}{H\sqrt{m}}\mathbf{J}_i^{(k)}(k)\mathbf{W}^{(k)}(k)\right\|_2. \end{aligned}$$

We have bounded the RHS in the proof for Lemma 3, thus

$$\left\|\mathcal{L}'^{(h)}(\theta(k))\right\|_F \leq \lambda_0 Q'(k).$$

Let $\theta(k,s)=\theta(k)-s\mathcal{L}'(\theta(k))$, we have

$$\begin{aligned}&\left\|u_i'^{(h)}(\theta(k))-u_i'^{(h)}(\theta(k,s))\right\|_F\\ =&\frac{\alpha_h}{H\sqrt{m}}\|x_i^{(h-1)}(k)a(k)^\top\prod_{l=h+1}^{H}\left(I+\frac{\alpha_h}{H\sqrt{m}}\mathbf{J}_i^{(l)}(k)\mathbf{W}^{(l)}(k)\right)\mathbf{J}_i^{(h)}(k)\\ &-x_i^{(h-1)}(k,s)a(k,s)^\top\prod_{l=h+1}^{H}\left(I+\frac{\alpha_h}{H\sqrt{m}}\mathbf{J}_i^{(l)}(k,s)\mathbf{W}^{(l)}(k,s)\right)\mathbf{J}_i^{(h)}(k,s)\|_F.\end{aligned}$$

Through standard calculations, we have

$$\begin{aligned}\|\mathbf{W}^{(l)}(k)-\mathbf{W}^{(l)}(k,s)\|_F\le&\eta Q'(k),\\ \|a(k)-a(k,s)\|_F\le&\eta Q'(k),\\ \|x_i^{(h-1)}(k)-x_i^{(h-1)}(k,s)\|_F\le&\eta c_x\frac{Q'(k)}{\sqrt{m}},\\ \|\mathbf{J}^{(l)}(k)-\mathbf{J}^{(l)}(k,s)\|_F\le&2(c_{x,0}+c_{w,0}c_x)\eta\beta Q'(k),\end{aligned}$$

where $c_x\triangleq(\sqrt{c_\sigma}L+\frac{c_{x,0}}{c_{w,0}})e^{3c_{w,0}L}$.

Given a set of matrices $\{\mathbf{A}_i,\mathbf{B}_i:i\in[n]\}$, if $\|\mathbf{A}_i\|_2\le M_i$, $\|\mathbf{B}_i\|_2\le M_i$ and $\|\mathbf{A}_i-\mathbf{B}_i\|_F\le\alpha_i M_i$, we have

$$\|\prod_{i=1}^{n}\mathbf{A}_i-\prod_{i=1}^{n}\mathbf{B}_i\|_F\le(\sum_{i=1}^{n}\alpha_i)\prod_{i=1}^{n}M_i.$$

So we have

$$\begin{aligned}&\|u_i'^{(h)}(\theta(k))-u_i'^{(h)}(\theta(k,s))\|_F\le\frac{4}{H}\alpha_h c_{x,0}La_{2,0}e^{2Lc_{w,0}}\eta\frac{Q'(k)}{\sqrt{m}}\left(\frac{c_x}{c_{x,0}}+\frac{2}{L}(c_{x,0}+c_{w,0}c_x)\beta\sqrt{m}\right)\\ &+\frac{4}{H}\alpha_h c_{x,0}La_{2,0}e^{2Lc_{w,0}}\eta\frac{Q'(k)}{\sqrt{m}}\left(4c_{w,0}(c_{x,0}+c_{w,0}c_x)\beta+L+1\right)\le\frac{32}{H}\alpha_h c_{x,0}a_{2,0}e^{2Lc_{w,0}}(c_{x,0}+c_{w,0}c_x)\beta\eta Q'(k).\end{aligned}$$

Thus we have

$$|I_2^i|\le 32c_{x,0}a_{2,0}e^{2Lc_{w,0}}(c_{x,0}+c_{w,0}c_x)\beta\eta^2Q'(k)^2=O(\eta^2\|y-u(k)\|_2),$$

where we use the bound on $\eta$ and that $\|y-u(0)\|_2=O(\sqrt{n})$. ■

*Proof of Lemma 7:*

$$I_1^i=-\eta\left\langle\mathcal{L}'(\theta(k)),u_i'(\theta(k))\right\rangle=-\eta\sum_{j=1}^{n}(u_j-y_j)\left\langle u_j'(\theta(k)),u_i'(\theta(k))\right\rangle\triangleq-\eta\sum_{j=1}^{n}(u_j-y_j)\sum_{h=1}^{H+1}\mathbf{G}_{ij}^{(h)}(k).$$

According to Section IV, we will only look at $\mathbf{G}^{(H)}$ matrix which has the following form $\mathbf{G}_{i,j}^{(H)}(k)=(x_i^{(H-1)}(k))^\top x_j^{(H-1)}(k)\cdot\frac{\alpha_H^2}{H^2m}\sum_{r=1}^{m}a_r^2\sigma'\left((\theta_r^{(H)}(k))^\top x_i^{(H-1)}(k)\right)\cdot\sigma'\left((\theta_r^{(H)}(k))^\top x_j^{(H-1)}(k)\right)$. Now we analyze $I_1(k)$. We can write $I_1$ in a more compact form with $\mathbf{G}(k)$:

$$I_1(k)=-\eta\mathbf{G}(k)(u(k)-y).$$

Now observe that

$$(y-u(k))^\top I_1(k)=\eta(y-u(k))^\top\mathbf{G}(k)(y-u(k))\ge\lambda_{\min}(\mathbf{G}(k))\|y-u(k)\|_2^2\ge\lambda_{\min}(\mathbf{G}^{(H)}(k))\|y-u(k)\|_2^2.$$

■

*Proof of Lemma 8:*

$$\|u(k+1)-u(k)\|_2^2 = \sum_{i=1}^n \Big(a(k+1)^\top x_i^{(H)}(k+1) - a(k)^\top x_i^{(H)}(k)\Big)^2$$

$$= \sum_{i=1}^n \Big([a(k+1)-a(k)]^\top x_i^{(H)}(k+1) + a(k)^\top [x_i^{(H)}(k+1) - x_i^{(H)}(k)]\Big)^2$$

$$\leq 2\|a(k+1)-a(k)\|_2^2 \sum_{i=1}^n \|x_i^{(H)}(k+1)\|_2^2 + 2\|a(k)\|_2^2 \sum_{i=1}^n \|x_i^{(H)}(k+1) - x_i^{(H)}(k)\|_2^2$$

$$\leq 8n\eta^2 c_{x,0}^2 Q'(k)^2 + 4n(\eta a_{2,0} c_x Q'(k))^2 = O(\eta^2 \|y-u(k)\|_2^2).$$

■

## References

[1] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, “Gradient descent finds global minima of deep neural networks,” in International conference on machine learning. PMLR, 2019, pp. 1675–1685.

[2] J. Sun, *Matrix perturbation analysis*, 2001, vol. 6.

[3] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, 2010.