A PROOFS OF BASIC PROPERTIES OF SEMI-NORM

Proof of Lemma 2.1. The first property is a direct consequence of the definition of the projection matrix P_X.

Notice that

$$\frac{1}{\varepsilon}(L(W+\varepsilon\Delta W)-L(W))=\frac{1}{\varepsilon}(L(WP_X+\varepsilon\Delta WP_X)-L(WP_X)).$$

By letting $\varepsilon\to 0$, the definition of the directional derivative implies

$$\langle\nabla L(W),\Delta W\rangle_F=\langle\nabla L(WP_X),\Delta WP_X\rangle_F=\langle\nabla L(WP_X)P_X,\Delta W\rangle_F,\forall\Delta W\in\mathbb{R}^{n_y\times n_x},$$

since $P_X=P_X^T$. This completes the proof of the second property.

The third property is obtained based on the fact that the orthogonal projection matrix satisfies $P_X=P_X^T=P_X^2=P_X^3$, since

$$\begin{aligned}&\langle\nabla L(W),V\rangle_F=\langle\nabla L(WP_X)P_X,V\rangle_F\\=&\langle\nabla L(WP_X)P_X^2,VP_X\rangle_F=\langle\nabla L(WP_X)P_X,V\rangle_X=\langle\nabla L(W),V\rangle_X.\end{aligned}$$

Set $V=\nabla L(W)$. Then the fourth property is implied by the third property.

For the last property, first recall that $\|W\|_X=\|WP_X\|_F$ and $P_X=X(X^TX)^\dagger X^T$. X is of full row rank if and only if P_X is identity matrix, which completes the proof.

□

Proof of Lemma 2.2. Because X is not full row rank, we know that $I-P_X\neq 0$. There exists W such that $W(I-P_X)\neq 0$. Applying the first property in Lemma 2.1, we have

$$L(\frac{1}{2}W+\frac{1}{2}WP_X)=L((\frac{1}{2}W+\frac{1}{2}WP_X)P_X)=L(WP_X)=\frac{1}{2}L(W)+\frac{1}{2}L(WP_X),$$

provided $W\neq WP_X$.

Hence, L is not strictly convex, which implies L is not strongly convex.

To prove the second property, it suffices to show that $g(W)=L(W)-\frac{\alpha(l)\lambda_{min}(XX^T)}{m}\|W\|_X^2$ is convex. It is obvious that

$$g(W)=L(W)-\frac{\alpha(l)}{m}\sum_{i=1}^m\|Wx_i-y_i\|_2^2+\frac{\alpha(l)}{m}(\|WX-Y\|_F^2-\lambda_{min}(X^TX)\|W\|_X^2).\quad(14)$$

$L(W)-\frac{\alpha(l)}{m}\sum_{i=1}^m\|Wx_i,y_i\|_F^2$ is convex, since $l(\cdot,y_i)$ is strongly convex. The Hessian of $\|WX-Y\|_F^2-\lambda_{min}(W^TW)\|WP_X\|_F^2$ has no negative eigenvalue, thus the second term in (14) is also convex. This completes the proof. □

B THE EXACT STATEMENTS OF THE MAIN THEOREMS

Define some quantities as follows:

$$q=\begin{cases}1-\alpha\eta_*(2-\eta_*\alpha), & 0<\eta_*\leq\frac{2}{\alpha+\beta}\\ 1-\beta\eta_*(2-\eta_*\beta), & \frac{2}{(\alpha+\beta)}<\eta_*<\frac{2}{\beta},\end{cases}\quad(15)$$

$$
\begin{aligned}
B_\delta &= \left(\frac{2\cdot rank(X)}{\delta} + \|W_*\|_X^2\right), \\
C_1 &= n_N\kappa^2 B_\delta \frac{C_0}{(\eta_0-\eta)^2/\eta_0^2} + \ln N, \\
C_2 &= n_N\kappa^2 B_\delta C_0 + \ln N, \\
C_3 &= n_N\kappa^2 B_\delta \frac{C_0}{(\eta_0-\eta)^2/\eta_0^2} + C_0\ln(\underline{N}), \\
C_4 &= n_N\kappa^2 B_\delta \frac{1}{(\eta_0-\eta)^2/\eta_0^2}, \\
C_5 &= n_N\kappa^2 B_\delta C_0 + C_0\ln(\underline{N}), \\
C_6 &= n_N\kappa^2 B_\delta,
\end{aligned}
$$

where $\underline{N}$ denotes the number of distinct elements in the set $\{n_1,\cdots,n_{N-1}\}$, $\eta_1 = \frac{2n_N}{N\beta}$, and $\eta_0 = \frac{2n_N}{e^{2c}N\beta}$ with $c>0$.

Theorem B.1. *Given any $c>0$, and $0<\delta<1/2$, define $\eta_0 = \frac{2n_N}{e^{2c}N\beta}$, and consider the learning rate $\eta<\eta_0$. There exists a constant $C := C(c)$, such that if*

$$
n_{min} \geq C\cdot C_1\cdot N, \tag{16}
$$

then with probability at least $1-\delta$ over the random Gaussian initialization, we have

$$
\mathcal{E}_{DLN}(t) \leq \left(1 - 4e^{-c\frac{\frac{\eta}{\eta_0}(1-\frac{\eta}{\eta_0})}{\kappa}}\right)^t \mathcal{E}_{DLN}(0).
$$

Theorem B.2. *Given any $c>0$, and $0<\delta<1/2$, define $\eta_0 = \frac{2n_N}{e^{2c}\beta N}$, and consider the learning rate $\eta<\eta_0$. There exists a constant $C := C(c)$, such that if*

$$
n_{min} \geq C\cdot C_3, \tag{17}
$$

then with probability at least $1-\delta$ over the random one peak projections and embeddings initialization, we have

$$
\mathcal{E}_{DLN}(t) \leq \left(1 - 4e^{-c\frac{\frac{\eta}{\eta_0}(1-\frac{\eta}{\eta_0})}{\kappa}}\right)^t \mathcal{E}_{DLN}(0).
$$

Specially, if $n_1 = n_2 = \cdots = n_{N-1} = n \geq \min\{n_N, n_0\}$, then the requirement (17) can be replaced by

$$
n \geq C\cdot C_4. \tag{18}
$$

Remark 7. Assume $L(a_N W_N\cdots W_1) = \frac{1}{2}\|a_N W_N\cdots W_1 X - Y\|_F^2$, and $n_1 = \cdots = n_{N-1} = n$. Then for Gaussian initialization, our Theorem B.1 leads to Theorem 4.1 in Du & Hu (2019). Similarly, for orthogonal initialization, our Theorem B.2 leads to Theorem 4.1 in Hu et al. (2020).

Next, we present a version of the theorem related to balanced initialization.

Theorem B.3. *Assume $n_1 = \cdots = n_{N-1} = n$. Given any $c>0$, and $0<\delta<1/2$, define $\eta_0 = \frac{2n_N}{e^{2c}\beta N}$, and consider the learning rate $\eta<\eta_0$. There exists a constant $C := C(c)$, such that as long as*

$$
n \geq C\cdot C_4. \tag{19}
$$

then with probability at least $1-\delta$ over special balanced initial, we have

$$
\mathcal{E}_{DLN}(t) \leq \left(1 - 4e^{-c\frac{\frac{\eta}{\eta_0}(1-\frac{\eta}{\eta_0})}{\kappa}}\right)^t \mathcal{E}_{DLN}(0).
$$

C INEQUALITIES IN CONVEX OPTIMIZATION

Convex optimization has been studied for about a century. Recall the definitions and basic inequalities for $\alpha-$strongly convex and $\beta-$Lipschitz functions.

Definition C.1. A continues differentiable function f is said to be $\beta-$ Lipschitz if the gradient ∇f is $\beta-$ Lipschitz, that is if for all x, y,

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\|, \tag{20}$$

f is said to be $\alpha-$strongly convex if for all x, y, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{21}$$

Proposition C.1. *If f is $\alpha-$strongly convex and ∇f is $\beta-$Lipschitz with respect to a (semi-)norm, then $\alpha \leq \beta$ and*

$$\langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \leq f(y) - f(x) \leq \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2, \tag{22}$$

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x - y\|^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|^2, \tag{23}$$

$$\|\nabla f(x) - \nabla f(y)\| \geq \alpha\|x - y\|, \tag{24}$$

$$f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2. \tag{25}$$

Proof of Proposition C.1. We only proof the last inequality.
Let $z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x))$. Since f is convex $\beta-$Lipschitz, we have

$$f(z) - f(x) \geq \langle \nabla f(x), z - x\rangle$$

and

$$f(z) - f(y) \leq \langle \nabla f(y), z - y\rangle + \frac{\beta}{2}\|z - y\|^2.$$

Thus,

$$\begin{aligned} f(x) - f(y) =& f(x) - f(z) + f(z) - f(y) \\ \leq& \langle \nabla f(x), x - z\rangle + \langle \nabla f(y), z - y\rangle + \frac{\beta}{2}\|z - y\|^2 \\ =& \langle \nabla f(x), x - y\rangle - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2. \end{aligned}$$

□

Before we start to prove Lemma D.1, let us first include and prove the following result.

Lemma C.2. ***1.*** *Assume L is $\alpha-$strongly convex, $\alpha > 0$. Denote a global minimizer of L by W_*. Then for any W,*

$$L(W_*) - L(W) \geq -\frac{1}{2\alpha}\|\nabla L(W)\|_X^2. \tag{26}$$

2. *Assume ∇L is $\beta-$Lipschitz, then*

$$L(W_*) - L(W) \leq -\frac{1}{2\beta}\|\nabla L(W)\|_X^2. \tag{27}$$

Proof of Lemma C.2. **1.** First, we know that $\nabla L(W_*) = 0$. L is $\alpha-$strongly convex, which implies the inequality (22) holds. Thus

$$L(V) - L(W) \geq \langle \nabla L(W), V - W\rangle_X + \frac{\alpha}{2}\|V - W\|_X^2 =: g(V).$$

Minimizing both sides in terms of V gives (26).

Now we focus on minimizing $g(V)$. Since $g(V) \in C^1$ and the global minimizer exits, we have

$$\nabla g(V^*) = \nabla L(W) P_X + \alpha(V^* - W) P_X = 0,$$

where V^* is a global minimizer for $g(V)$. Thus,

$$g(V^*) = -\frac{1}{2\alpha} \|\nabla L(W)\|_X^2 . \tag{28}$$

2. Applying proposition C.1 to a $\beta-$Lipschitz function ∇L, we obtain

$$\begin{aligned} & L(W_*) - L(W) \\ \leq & \langle \nabla L(W_*), W_* - W \rangle_X - \frac{1}{2\beta} \|\nabla L(W) - \nabla L(W_*)\|_X^2 \\ = & -\frac{1}{2\beta} \|\nabla L(W)\|_X^2 . \end{aligned}$$

□

D CONVERGENCE REGION

In this section, we study a class of the convergence region for deep linear neural networks, which works for deterministic initialization. Define $A|_{\mathcal{R}(X)} = AX^T(XX^T)^- X = AP_X$, and view $A|_{\mathcal{R}(X)}$ as a linear operator on $\mathcal{R}(X)$.

Recall the optimization problem

$$\underset{W_1,\cdots,W_N}{\text{minimize}} \quad L^N(W_1, \cdots W_N) := \frac{1}{m} \sum_{i=1}^{m} l(a_N W_{N:1} x_i, y_i) = L(a_N W_{N:1}), \tag{29}$$

and GD

$$\begin{cases} W_j(t+1) = W_j(t) - \eta \frac{\partial L^N}{\partial W_j}(W_1(t), \cdots, W_N(t)), j = 1, \cdots, N, \\ \text{where } \frac{\partial L^N}{\partial W_j}(W_1, \cdots, W_N) = a_N (W_{N:j+1})^T \nabla L(a_N W_{N:1}) (W_{j-1:1})^T, \end{cases} \tag{30}$$

where the normalization factor $a_N = \frac{1}{\sqrt{n_1 n_2 \cdots n_{N-1} n_N}}$.

The following theorem generalizes the idea from the recent work (Du & Hu, 2019; Hu et al., 2020).

For notational convenience, we denote $W_{j:i}(t) = W_j(t) \cdots W_i(t)$, $L_t = L(a_N W_{N:1}(t))$, $\nabla L_t = \nabla L(a_N W_{N:1}(t))$ etc.

Lemma D.1. *Assume the initialization satisfies the following conditions simultaneously:*

$$\begin{cases} \sigma_{max}(W_{N:i+1}(0)) \leq e^{c_1/2} (n_{N-1:i})^{1/2}, 1 \leq i \leq N-1, \\ \sigma_{min}(W_{N:i+1}(0)) \geq e^{-c_2/2} (n_{N-1:i})^{1/2}, 1 \leq i \leq N-1, \\ \sigma_{max}(W_{i-1:1}(0)|_{\mathcal{R}(X)}) \leq e^{c_1/2} (n_{i-1:1})^{1/2}, 2 \leq i \leq N, \\ \sigma_{min}(W_{i-1:1}(0)|_{\mathcal{R}(X)}) \geq e^{-c_2/2} (n_{i-1:1})^{1/2}, 2 \leq i \leq N, \\ \|W_{j:i}(0)\| \leq M/2 \cdot N^\theta (\prod_{i \leq k \leq j-1} n_k \cdot \max\{n_{i-1}, n_j\})^{1/2}, 1 < i \leq j < N, \\ L_0 - L(W_*) \leq \beta B_0 =: B, \end{cases} \tag{31}$$

where c_1, c_2, M are positive constant and $\theta \geq 0$.
Notice that B_0 is a proper upper bound for $\|a_N W_{N:1}(0)\|_X^2 + \|W_*\|_X^2$.

Set the learning rate $\eta = \frac{(1-\varepsilon) 2 n_N}{e^{6c_1 + 3c_2} \beta N}$, where $0 < \varepsilon < 1$. Define $\gamma = \frac{2 e^{6c_1} \varepsilon \alpha N}{n_N}$.

Assume that

$$n_{min} \geq \frac{C(c_1, c_2) M^2 \kappa^2 B_0}{\varepsilon^2} N^{2\theta} n_N. \tag{32}$$

Then the GD (30) satisfies

$$L_t - L(W_*) \leq (1 - \eta\gamma)^t (L_0 - L(W_*)), t = 1, 2, \cdots .$$

Definition D.1. For given $c_1, c_2, M, B_0 > 0$, and $\theta \geq 0$, we define the convergence region $\mathcal{R}(c_1, c_2, \theta, M, B_0)$ by the set of initialization that satisfies the inequality system (31).

Remark 8. The condition (31) describes the convergence region for initialization and the condition (32) describes the overparameterization for deep linear neural networks. At this time, it is not clear how large this convergence region is. Later, we will show that the properly scaled random initialization with some extra mild overparameterization conditions will fall into this convergence region with high probability.

Proof of Lemme D.1. To prove Lemma D.1, it suffices to show that the following three properties hold $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ for all $t = 0, 1, \cdots$.

1. $\mathcal{A}(t)$:
$$L_t - L(W_*) \leq (1 - \eta\gamma)^t (L_0 - L(W_*)).$$

2. $\mathcal{B}(t)$:
$$\begin{cases} \sigma_{max}(W_{N:i+1}(t)) \leq e^{c_1}(n_{N-1:i})^{1/2}, 1 \leq i \leq N-1, \\ \sigma_{min}(W_{N:i+1}(t)) \geq e^{-c_2}(n_{N-1:i})^{1/2}, 1 \leq i \leq N-1, \\ \sigma_{max}(W_{i-1:1}(t)|_{\mathcal{R}(X)}) \leq e^{c_1}(n_{i-1:1})^{1/2}, 2 \leq i \leq N, \\ \sigma_{min}(W_{i-1:1}(t)|_{\mathcal{R}(X)}) \geq e^{-c_2}(n_{i-1:1})^{1/2}, 2 \leq i \leq N, \\ \|W_{j:i}(t)\| \leq M \cdot N^{\theta}(\frac{1}{n_{min}} \prod_{i-1 \leq k \leq j} n_k)^{1/2}, 1 < i \leq j < N. \end{cases}$$

3. $\mathcal{C}(t)$:
$$\|W_i(t) - W_i(0)\|_F \leq \frac{2e^{2c_1}\sqrt{2\beta B}}{\sqrt{n_N}\gamma} =: R, 1 \leq i \leq N.$$

Using simultaneous induction, the proof of Lemma D.1 is divided into the following 3 claims.

Claim 1. $\mathcal{A}(0), \cdots, \mathcal{A}(t), \mathcal{B}(0), \cdots, \mathcal{B}(t) \implies \mathcal{C}(t+1)$.

Claim 2. $\mathcal{C}(t) \implies \mathcal{B}(t)$, if $n_{min} \geq \frac{C(c_1,c_2)M^2\kappa^2 B_0}{\varepsilon^2} N^{2\theta} n_N$, where $C(c_1, c_2)$ is a positive constant only depend on c_1, c_2.

Claim 3. $\mathcal{A}(t), \mathcal{B}(t) \implies \mathcal{A}(t+1)$, if $n_{min} \geq C(c_1, c_2)M^2 B_0 N^{2\theta} n_N$, where $C(c_1, c_2)$ is a positive constant only depend on c_1, c_2.

□

Proof of Claim 1. As a consequence of Lemma C.2 and Lemma 2.1, and $\mathcal{A}(s)$, $s \leq t$, we have

$$\begin{aligned} &\|\nabla L(a_N W_{N:1}(s))\|_F^2 = \|\nabla L_s - \nabla L(W_* P_X)\|_X^2 \\ \leq &2\beta[L_s - L(W_*)] \\ \leq &2\beta (1 - \eta\gamma)^s B. \end{aligned} \tag{33}$$

From $\mathcal{A}(0), \cdots, \mathcal{A}(t), \mathcal{B}(0), \cdots, \mathcal{B}(t)$, we have for any $0 \leq s \leq t$,

$$\begin{aligned} \left\| \frac{\partial L}{\partial W_i}(s) \right\|_F &\leq a_N \|W_{N:i+1}(s)\| \|\nabla L(a_N W_{N:1}(s))\|_F \left\|W_{i-1:1}(s)|_{\mathcal{R}(X)}\right\| \\ &\leq \frac{e^{2c_1}}{\sqrt{n_N}} \|\nabla L(a_N W_{N:1}(s))\|_F \\ &\leq \frac{e^{2c_1}}{\sqrt{n_N}} \sqrt{2\beta (1 - \eta\gamma)^s B}. \end{aligned} \tag{34}$$

Then,

$$\begin{aligned}\|W_i(t+1)-W_i(0)\|_F &\le \sum_{s=0}^{t}\|W_i(s+1)-W_i(s)\|_F\\ &= \sum_{s=0}^{t}\left\|\eta\frac{\partial L}{\partial W_i}(s)\right\|_F\\ &\le \eta\frac{e^{2c_1}}{\sqrt{n_N}}\sqrt{2\beta B}\sum_{s=0}^{t}(1-\eta\gamma)^{s/2}\\ &\le \eta\frac{e^{2c_1}}{\sqrt{n_N}}\sqrt{2\beta B}\sum_{s=0}^{t}(1-\eta\gamma/2)^{s}\\ &\le \frac{2e^{2c_1}\sqrt{2\beta B}}{\sqrt{n_N}\gamma}=R.\end{aligned}$$

This proves $\mathcal{C}(t+1)$. □

Proof of Claim 2. Let $\delta_i = W_i(t)-W_i(0), 1\le i\le N$. Using $\mathcal{C}(t)$, we have $\|\delta_i\|_F \le R, 1\le i\le N$. Set $\varepsilon_1 = e^{-c_1/2}\min\{e^{c_1}-e^{c_1/2}, e^{-c_2/2}-e^{-c_2}, 1/2\}$.
It is suffices to show that

$$\|W_{N:i}(t)-W_{N:i}(0)\| \le e^{c_1/2}\varepsilon_1(n_{N-1}n_{N-1}\cdots n_{i-1})^{1/2}, 1<i\le N, \tag{35}$$

$$\left\|(W_{i:1}(t)-W_{i:1}(0))|_{\mathcal{R}(X)}\right\| \le e^{c_1/2}\varepsilon_1(n_1n_2\cdots n_{i-1})^{1/2}, 1\le i<N, \tag{36}$$

and

$$\|W_{j:i}(t)-W_{j:i}(0)\| \le M/2\cdot N^{\theta}\left(\frac{1}{n_{min}}\prod_{i-1\le k\le j}n_k\right)^{1/2}, 1<i\le j<N, \tag{37}$$

because $\sigma_{min}(A+B)\ge\sigma_{min}(A)-\sigma_{max}(B)=\sigma_{min}(A)-\|B\|$ and $\sigma_{max}(A+B)\le\sigma_{max}(A)+\sigma_{max}(B)=\|A\|+\|B\|$ (e.g. see Theorem 1.3 in Chafaı et al. (2009)).
Case 1. We first prove (37).
For $1\le i<j\le N$, we can write $W_{j:i}(t)=(W_j(0)+\delta_j)\cdots(W_i(0)+\delta_i)$.
Expanding the above product, each term has the form:

$$W_{j:(k_s+1)}(0)\cdot\delta_{k_s}\cdot W_{(k_s-1):(k_{s-1}+1)}(0)\cdot\delta_{k_{s-1}}\cdots\delta_{k_1}\cdot W_{(k_1-1):i}(0), \tag{38}$$

where $i\le k_1<\cdots<k_s\le j$ are positions at which perturbation terms δ_{k_l} are taken out.
Notice that the convergence region assumptions (31) implies that for any $1<i\le j<N$,

$$\|W_{j:i}(0)\| \le M/2\cdot N^{\theta}\left(\prod_{i\le k\le j-1}n_k\cdot\max\{n_{i-1},n_j\}\right)^{1/2} \le M\cdot N^{\theta}\left(\frac{\prod_{i-1\le k\le j}n_k}{n_{min}}\right)^{1/2}. \tag{39}$$

WLOG, assume $M\ge 1$. If $i=j+1$, then

$$\|W_{j:i}(0)\| = \|I\| \le M\cdot N^{\theta}(n_j/n_{min})^{1/2}.$$

Assume $i>1, j<N$, applying inequality (39) as well as the following inequality

$$\sum_{s=1}^{j-i+1}\binom{j-i+1}{s}x^s = (1+x)^{j-i+1}-1 \le (1+x)^N-1, \forall x\ge 0,$$

we obtain that

$$\begin{aligned}&\|W_{j:i}(t)-W_{j:i}(0)\|\\ \le&\sum_{s=1}^{j-i+1}\binom{j-i+1}{s}R^s(M\cdot N^{\theta})^{s+1}n_{min}^{-s/2}(n_{i-1}\cdots n_j/n_{min})^{1/2}\\ \le&M\cdot N^{\theta}(n_{i-1}\cdots n_j/n_{min})^{1/2}[(1+R\cdot M\cdot N^{\theta}/\sqrt{n_{min}})^N-1]\\ \le&\varepsilon_1 M\cdot N^{\theta}(n_{i-1}\cdots n_j/n_{min})^{1/2}.\end{aligned}$$

The last line holds due to the following reasons:
there exists absolute constant $A_1, A_2 > 0$ such that

$$(1+x)^N - 1 \le A_2 x N,$$

if $x \ge 0$, $N \ge 1$, and $xN \le A_1$. Since there exists positive constant $C(c_1, c_2)$, which only depends on c_1, c_2, such that when

$$n_{min} \ge \frac{C(c_1,c_2) M^2 \kappa^2 B_0}{\varepsilon^2} N^{2\theta} n_N \tag{40}$$

we can have

$$R \cdot M \cdot N^{\theta+1}/\sqrt{n_{min}} \le A_1,$$

as well as

$$[(1 + R \cdot M \cdot N^\theta/\sqrt{n_{min}})^N - 1] \le A_2 \cdot M \cdot R \cdot N^{\theta+1}/\sqrt{n_{min}} \le \varepsilon_1 = \varepsilon_1(c_1, c_2).$$

Case 2. The proof of (35) is similar. Set $j = N$, we can save the factor $M \cdot N^\theta$ from previous calculation, which means

$$\begin{aligned}
&\|W_{N:i}(t) - W_{N:i}(0)\| \\
\le& e^{c_1/2} \sum_{s=1}^{N-i+1} \binom{N-i+1}{s} R^s (M \cdot N^\theta)^s n_{min}^{-s/2} (n_{i-1} \cdots n_{N-1})^{1/2} \\
\le& e^{c_1/2} (n_{i-1} \cdots n_{N-1})^{1/2} [(1 + R \cdot M \cdot N^\theta/\sqrt{n_{min}})^N - 1] \\
\le& e^{c_1/2} \varepsilon_1 (n_{i-1} \cdots n_{N-1})^{1/2}, i \ge 2,
\end{aligned}$$

where the last line is implied by equation (40).
Case 3. Similarly, we have

$$\begin{aligned}
&\left\|W_{j:1}(t)|_{\mathcal{R}(X)} - W_{j:1}(0)|_{\mathcal{R}(X)}\right\| \\
\le& e^{c_1/2} \sum_{s=1}^{j} \binom{j}{s} R^s (M \cdot N^\theta)^s n_{min}^{-s/2} (n_1 \cdots n_j)^{1/2} \\
\le& e^{c_1/2} (n_1 \cdots n_j)^{1/2} [(1 + R \cdot M \cdot N^\theta/\sqrt{n_{min}})^N - 1] \\
\le& e^{c_1/2} \varepsilon_1 (n_1 \cdots n_j)^{1/2}, j \le N-1
\end{aligned}$$

This proves $\mathcal{B}(t)$.

□

Proof of Claim 3. The GD (7) implies

$$\begin{aligned}
&W_{N:1}(t+1) \\
=& \left(W_N(t) - \eta \frac{\partial L^N}{\partial W_N}(t)\right)\left(W_{N-1}(t) - \eta \frac{\partial L^N}{\partial W_{N-1}}(t)\right) \cdots \left(W_1(t) - \eta \frac{\partial L^N}{\partial W_1}(t)\right) \\
=& W_{N:1}(t) - \eta \cdot a_N \sum_{i=1}^{N} W_{N:i+1}(t) W_{N:i+1}^T(t) \nabla L(a_N W_{N:1}(t)) (W_{i-1:1}(t))^T (W_{i-1:1}(t)) + E(t),
\end{aligned}$$

where $E(t)$ contains all high-order terms (those with η^2 or higher). Define a linear operator

$$P(t)[A] = a_N^2 \sum_{i=1}^{N} W_{N:i+1}(t) W_{N:i+1}^T(t) (A P_X) (W_{i-1:1}(t)|_{\mathcal{R}(X)})^T W_{i-1:1}(t)|_{\mathcal{R}(X)}, \tag{41}$$

for any $A \in \mathbb{R}^{n_N \times n_0}$.

Now we have

$$a_N W_{N:1}(t+1) = a_N W_{N:1}(t) - \eta \cdot P(t)[\nabla L(a_N W_{N:1}(t) P_X)] + a_N E(t). \tag{42}$$

Easy to check that $P(t)[\cdot]$ is a sum of positive semidefinite linear operator.

The following proposition describes the eigenvalues of the linear operator $P(t)[\cdot]$.

Proposition D.2. *Let S_1, S_2 be symmetric matrices. Suppose $S_1 = U\Lambda_1 U^T$, $S_2 = V\Lambda_2 V^T$, where $U = [u_1, u_2, \cdots, u_m]$, and $V = [v_1, v_2, \cdots, v_n]$ are othogonal matrices, and $\Lambda_1 = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$ and $\Lambda_2 = diag(\mu_1, \mu_2, \cdots, \mu_n)$ are diagonal matrices. Then the linear operator $L(A) := S_1 A S_2$ is orthogonally diagonalizable, and $L(A_{ij}) = \lambda_i \mu_j A_{ij}$, where $\lambda_i \mu_j$ represent all eigenvalues corresponding to their eigenvectors $A_{ij} = u_i v_j^T$.*

Applying this proposition and the assumption $\mathcal{B}(t)$, we obtain the upper bound and lower bound for the maximum and minimum eigenvalues of positive definite operator $P(t)$, respectively,

$$\lambda_{max}(P(t)) \leq a_N^2 \sum_{i=1}^{N} \sigma_{max}^2(W_{i-1:1}(t)|_{\mathcal{R}(X)}) \cdot \sigma_{max}^2(W_{N:i+1}(t)) \leq \frac{N}{n_N} e^{2c_1},$$

and

$$\lambda_{min}(P(t)) \geq a_N^2 \sum_{i=1}^{N} \sigma_{min}^2(W_{i-1:1}(t)|_{\mathcal{R}(X)}) \cdot \sigma_{min}^2(W_{N:i+1}(t)) \geq \frac{N}{n_N} e^{-2c_2}. \tag{43}$$

In conclusion, we have

$$\lambda_{max}(P(t)) \leq \frac{N}{n_N} e^{2c_1}, \text{and } \lambda_{min}(P(t)) \geq \frac{N}{n_N} e^{-2c_2}. \tag{44}$$

With learning rate $\eta = \eta_\varepsilon = \frac{(1-\varepsilon)2n_N}{e^{6c_1+3c_2}\beta N}$, $0 < \varepsilon < 1$, we have

$$\begin{aligned}
&L_{t+1} - L_t \\
&\leq \langle \nabla L_t, -\eta P(t)[\nabla L_t] \rangle_X + \langle \nabla L_t, a_N E(t) \rangle_X + \frac{\beta}{2} \left\| \eta P(t)[\nabla L_t] - a_N E(t) \right\|_X^2 \\
&= \langle \nabla L_t, -\eta P(t)[\nabla L_t] \rangle + \frac{\beta}{2} \eta^2 \left\| P(t)[\nabla L_t] \right\|_X^2 + F(t) \\
&\leq -\left(\eta \lambda_{min}(P(t)) - \frac{\beta}{2} \eta^2 \lambda_{max}^2(P(t)) \right) \left\| \nabla L_t \right\|_X^2 + F(t) \\
&\leq -e^{-2c_2} \frac{N}{n_N} \eta \left(1 - e^{4c_1+2c_2} \frac{\beta}{2} \eta \frac{N}{n_N} \right) \left\| \nabla L_t \right\|_X^2 + F(t),
\end{aligned} \tag{45}$$

where

$$F(t) = \langle \nabla L_t, a_N E(t) \rangle_X + \frac{\beta}{2} \left\| \eta P(t)[\nabla L_t] - a_N E(t) \right\|_X^2 - \frac{\beta}{2} \eta^2 \left\| P(t)[\nabla L_t] \right\|_X^2.$$

We claim that $F(t)$ is small enough, such that

$$\begin{aligned}
&L_{t+1} - L_t \\
&\leq -e^{-2c_2} \frac{N}{n_N} \eta \left(1 - e^{4c_1+2c_2} \frac{\beta}{2} \eta \frac{N}{n_N} \right) \left\| \nabla L_t \right\|_X^2 + F(t) \\
&\leq -e^{-3c_2} \frac{N}{n_N} \eta \left(1 - e^{6c_1+3c_2} \frac{\beta}{2} \eta \frac{N}{n_N} \right) \left\| \nabla L_t \right\|_X^2 \\
&= -e^{-6(c_1+c_2)} \frac{2\varepsilon(1-\varepsilon)}{\beta} \left\| \nabla L_t \right\|_X^2.
\end{aligned} \tag{46}$$

Assuming this claim for the moment, we complete the proof. Combining (26) and (46), we have

$$\begin{cases} L_{t+1} - L_t \leq -e^{-6(c_1+c_2)} \frac{2\varepsilon(1-\varepsilon)}{\beta} \left\| \nabla L_t \right\|_X^2, \\ L(W_*) - L_t \geq -\frac{1}{2\alpha} \left\| \nabla L_t \right\|_X^2, \end{cases}$$

which implies

$$L_{t+1} - L(W_*) \leq \left(1 - e^{-6(c_1+c_2)} \frac{4\varepsilon(1-\varepsilon)}{\kappa} \right) (L_t - L(W_*)), \tag{47}$$

that is

$$L_t - L(W_*) \leq \left(1 - e^{-6(c_1+c_2)} \frac{4\varepsilon(1-\varepsilon)}{\kappa}\right)^t (L_0 - L(W_*)) = (1-\eta\gamma)^t (L_0 - L(W_*)). \quad (48)$$

Estimate $F(t)$
Notice that

$$\begin{aligned}
&|F(t)| \\
\leq &\|\nabla L_t\|_X \|a_N E(t)\|_X + \frac{\beta}{2}(2\eta\lambda_{max}(P(t)) \|\nabla L_t\|_X \|a_N E(t)\|_X + \|a_N E(t)\|_X^2) \\
= &: I_1 + I_2.
\end{aligned}$$

From (34), we have

$$\left\| \frac{\partial L}{\partial W_i}(t) \right\|_F \leq \frac{e^{2c_1}}{\sqrt{n_N}} \|\nabla L(a_N W_{N:1}(t))\|_F = \frac{e^{2c_1}}{\sqrt{n_N}} \|\nabla L(a_N W_{N:1}(t))\|_X =: K.$$

Expanding the product

$$W_{N:1}(t+1) = \left(W_N(t) - \eta \frac{\partial L^N}{\partial W_N}(t)\right)\left(W_{N-1}(t) - \eta \frac{\partial L^N}{\partial W_{N-1}}(t)\right) \cdots \left(W_1(t) - \eta \frac{\partial L^N}{\partial W_1}(t)\right),$$

each term has the form:

$$\Delta = W_{N:(k_s+1)}(t) \cdot \eta \frac{\partial L}{\partial W_{k_s}}(t) \cdot W_{(k_s-1):(k_{s-1}+1)}(t) \cdot \eta \frac{\partial L}{\partial W_{k_{s-1}}}(t) \cdots \eta \frac{\partial L}{W_{k_1}}(t) \cdot W_{(k_1-1):1}(t),$$

where $1 \leq k_1 < k_2 < \cdots < k_s \leq N$.

As a direct consequence of inequality $\mathcal{B}(t)$ and inequality (39), we obtain

$$\|\Delta\|_X = \|\Delta P_X\|_F \leq \frac{1}{a_N \sqrt{n_N}} e^{2c_1} (\eta K)^s \left(\frac{M \cdot N^\theta}{\sqrt{n_{min}}}\right)^{s-1},$$

Recall that $E(t)$ contains all high-order terms (those with η^2 or higher) in the expansion of the product. Thus, $E(t)$ can be expressed as follows:

$$\sum_{s=2}^{N} \sum_{1 \leq k_1 < k_2 < \cdots < k_s \leq N} W_{N:(k_s+1)}(t) \cdot \eta \frac{\partial L}{\partial W_{k_s}}(t) \cdot W_{(k_s-1):(k_{s-1}+1)}(t) \cdot \eta \frac{\partial L}{\partial W_{k_{s-1}}}(t) \cdots \eta \frac{\partial L}{W_{k_1}}(t) \cdot W_{(k_1-1):1}(t).$$

Set $\xi = \min\{(e^{-2c_2} - e^{-3c_2})/e^{4c_1+1}, \frac{1}{4}(e^{6c_1} - e^{4c_1})/e^{6c_1+1}, \frac{1}{2}(e^{6c_1} - e^{4c_1})^{1/2}/e^{4c_1+1}, 1\}$.
Recall the inequality $\binom{N}{s} \leq (eN)^s$. Thus, we have

$$\begin{aligned}
&a_N \|E(t)\|_X \\
\leq &\frac{1}{\sqrt{n_N}} e^{2c_1} \sum_{s=2}^{N} \binom{N}{s} (\eta K)^s \left(\frac{M \cdot N^\theta}{\sqrt{n_{min}}}\right)^{s-1} \\
\leq &\frac{1}{\sqrt{n_N}} \left(\frac{M \cdot N^\theta}{\sqrt{n_{min}}}\right)^{-1} e^{2c_1} \sum_{s=2}^{N} (eN)^s (\eta K)^s \left(\frac{M \cdot N^\theta}{\sqrt{n_{min}}}\right)^s \\
\leq &\frac{1}{\sqrt{n_N}} e^{2c_1} (\eta e K N) \frac{\eta e K M \cdot N^{\theta+1}/\sqrt{n_{min}}}{1 - \eta e K M \cdot N^{\theta+1}/\sqrt{n_{min}}} \\
\leq &\xi \frac{N}{n_N} \eta \cdot e^{4c_1+1} \|\nabla L(a_N W_{N:1}(t))\|_X \ (\text{ if } \eta e K M \cdot N^{\theta+1}/\sqrt{n_{min}} < \xi/(1+\xi)) \\
= &\xi \cdot e^{4c_1+1} \left(\eta \frac{N}{n_N}\right) \|\nabla L(a_N W_{N:1}(t))\|_X .
\end{aligned} \quad (49)$$

Using (33) and the upper bound of η, we know that there exists constant $C(c_1, c_2)$, such that

$$n_{min} \geq C(c_1, c_2) M^2 \cdot B_0 N^{2\theta} n_N,$$

and

$$\eta eKM \cdot N^{\theta+1}/\sqrt{n_{min}} \leq \frac{2\sqrt{2}M \cdot e^{1+2c_1}\sqrt{B_0}N^{\theta}\sqrt{n_N}}{\sqrt{n_{min}}} = \frac{1}{C'(c_1,c_2)} \leq \frac{\xi}{2} \leq \frac{\xi}{1+\xi}.$$

Using (49), we have

$$I_1 \leq \xi \cdot e^{4c_1+1}\left(\eta\frac{N}{n_N}\right)\|\nabla L_t\|_X^2 \leq (e^{-2c_2}-e^{-3c_2})\left(\eta\frac{N}{n_N}\right)\|\nabla L_t\|_X^2, \tag{50}$$

and

$$\begin{aligned}
&I_2\\
\leq&\frac{\beta}{2}\left(2\xi\cdot e^{6c_1+1}\left(\eta^2\frac{N^2}{n_N^2}\right)\|\nabla L_t\|_X^2+\xi^2\cdot e^{8c_1+2}\left(\eta^2\frac{N^2}{n_N^2}\right)\|\nabla L_t\|_X^2\right)\\
\leq&(e^{6c_1}-e^{4c_1})\frac{\beta}{2}\eta^2\frac{N^2}{n_N^2}\|\nabla L_t\|_X^2.
\end{aligned}$$

Thus, (46) valid.
This proves $\mathcal{A}(t)$.

□

As a direct consequence of the proof Lemma D.1, we can obtain the following lemma.

Lemma D.3. *Assume all assumptions in Lemma D.1 hold. For any $\tau > 0$, we can choose new constants c_1, c_2 as well as $C := C(c_1, c_2)$ such that the overparameterization assumption (32) in Lemma D.1 hold and*

$$\|R(t)\|_X \leq \tau\|a_N W_{N:1}(t) - W_*\|_X, \tag{51}$$

where

$$a_N W_{N:1}(t+1) = a_N W_{N:1}(t) - \frac{N}{n_N}\eta\nabla L(a_N W_{N:1}(t)) + R(t).$$

Proof of Lemma D.3. Due to (33), (42), (44), (49), and lemma C.2, we have

$$\begin{aligned}
\|R(t)\|_X =& \left\|a_N E(t) + \eta\left(\frac{N}{n_N}\nabla L_t - P(t)[\nabla L_t]\right)\right\|_X\\
\leq& \|a_N E(t)\|_X + \eta\max\left\{\lambda_{max}(P(t)) - \frac{N}{n_N}, \frac{N}{n_N} - \lambda_{min}(P(t))\right\}\|\nabla L_t\|_X\\
\leq&(C'\cdot\xi + \max\{e^{2c_1}-1, 1-e^{-2c_2}\})\cdot\eta\frac{N}{n_N}\cdot\|\nabla L_t\|_X\\
\leq&\frac{2\sqrt{2\beta(L_t - L(W_*))}}{e^{6c_1+3c_2}\cdot\beta}\cdot(C'\cdot\xi + \max\{e^{2c_1}-1, 1-e^{-2c_2}\}).
\end{aligned}$$

Due to the fact that $L_t - L(W_*)$ is non-increasing in t, and C' is a constant only depend on c_1, c_2, we can choose small enough positive c_1, c_2 and ξ, which depends on τ, such that

$$\|R(t)\|_X \leq \tau\frac{\sqrt{2\beta(L_t - L(W_*))}}{\beta} \leq \tau\|a_N W_{N:1}(t) - W_*\|_X.$$

□

Lemma D.4. *Assume $\tau \in [0, 1)$. Consider a discrete dynamical system $V(t)$ such that,*

$$V(t+1) = V(t) - \eta_*\nabla L(V(t)) + R(t),$$

where $\|R(t)\|_X \leq \tau\|V(t) - W_*\|_X$. If $\eta_* \leq 2/\beta$, we have

$$\|V(t) - W_*\|_X^2 \leq (q+7\tau)^t\|V(0) - W_*\|_X^2,$$

where q is defined in (15).

Proof of Lemma D.4. Set $\Delta(t) = V(t) - W_*$ and $\tau' = \tau \left\|\Delta(t)\right\|_X$. Notice that

$$\Delta(t+1) = \Delta(t) - \eta_*(\nabla L(V(t)) - \nabla L(W_*)) + R(t),$$

and

$$\begin{aligned}
&\left\|\Delta(t+1)\right\|_X^2 \\
\leq&\eta_*^2 \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X^2 - 2\eta_* \langle \Delta(t), \nabla L(V(t)) - \nabla L(W_*)\rangle_X \\
&+ \left\|\Delta(t)\right\|_X^2 + (2\left\|\Delta(t)\right\|_X + 2\eta_* \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X + \tau')\tau'.
\end{aligned}$$

By inequality (23),

$$\begin{aligned}
&\left\|\Delta(t+1)\right\|_X^2 \\
\leq& \left\|\Delta(t)\right\|_X^2 - 2\eta_* \langle \Delta(t), \nabla L(V(t)) - \nabla L(W_*)\rangle_X \\
&+ \eta_*^2 \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X^2 + 7\tau \left\|\Delta(t)\right\|_X^2 \\
=&(1+7\tau) \left\|\Delta(t)\right\|_X^2 - 2\eta_* \langle \Delta(t), \nabla L(V(t)) - \nabla L(W_*)\rangle_X \\
&+ \eta_*^2 \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X^2 \\
\leq&(1+7\tau) \left\|\Delta(t)\right\|_X^2 - 2\eta_* \frac{\alpha\beta}{\alpha+\beta} \left\|\Delta(t)\right\|_X^2 \\
&+ \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right) \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X^2.
\end{aligned}$$

Case 1: $\frac{2}{\alpha+\beta} < \eta_* < \frac{2}{\beta}$.
In this case, we have

$$\begin{aligned}
&\left\|\Delta(t+1)\right\|_X^2 \\
\leq&(1+7\tau) \left\|\Delta(t)\right\|_X^2 - 2\eta_* \frac{\alpha\beta}{\alpha+\beta} \left\|\Delta(t)\right\|_X^2 + \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right) \left\|\nabla L(V(t)) - \nabla L(W_*)\right\|_X^2 \\
\leq&(1+7\tau) \left\|\Delta(t)\right\|_X^2 - 2\eta_* \frac{\alpha\beta}{\alpha+\beta} \left\|\Delta(t)\right\|_X^2 + \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right) \beta^2 \left\|\Delta(t)\right\|_X^2 \\
\leq& (1 + 7\tau - \beta\eta_*(2 - \eta_*\beta)) \left\|\Delta(t)\right\|_X^2 \\
=&(q+7\tau) \left\|\Delta(t)\right\|_X^2.
\end{aligned}$$

Case 2: $0 < \eta_* \leq \frac{2}{\alpha+\beta}$.
Similarly, we have

$$\left\|\Delta(t+1)\right\|_X^2 \leq (1 + 7\tau - \alpha\eta_*(2 - \eta_*\alpha)) \left\|\Delta(t)\right\|_X^2 = (q+7\tau) \left\|\Delta(t)\right\|_X^2.$$

In both cases, we have $\left\|\Delta(t+1)\right\|_X^2 \leq (q+7\tau) \left\|\Delta(t)\right\|_X^2$.

Thus, $\left\|\Delta(t)\right\|_X^2 \leq (q+7\tau)^t \left\|\Delta(0)\right\|_X^2$.

□

Next, we will show that the trajectories of the GD (30) for deep linear neural networks (29) are close to those of GD (2) for the corresponding convex problem (1).

Lemma D.5. *Consider the GD for the deep linear neural networks (30) with learning rate $\eta < \eta_1$ for $a_N W_{N:1}(t), t = 0, 1, \cdots$, and the GD (2) with learning rate $\eta_* = \frac{N}{n_N}\eta$ for $W(t), t = 0, 1, \cdots$.*

Assume $C(c_1, c_2)$ exists in Lemma D.1 for any $c_1, c_2 > 0$. For any $\tau \in (0, 1)$, $\eta < \eta_1$ (η_1 defined in B), we can choose $c_1, c_2 > 0$ and the constant $C = C(c_1, c_2) = C'(\tau, \eta/\eta_1)$, such that inequality (51) holds, given initialization condition (31), and overparameterization condition

$$n_{min} \geq CM^2\kappa^2 B_0 N^{2\theta} n_N. \tag{52}$$

Furthermore, we have

$$\|a_N W_{N:1}(t) - W(t)\|_X^2 \leq D(\tau, q, t) \|a_N W_{N:1}(0) - W_*\|_X^2 , \tag{53a}$$

$$|\mathcal{E}_{DLN}(t) - \mathcal{E}(t)| \leq \beta \left(q^{t/2} \sqrt{D(\tau, q, t)} + \frac{1}{2} D(\tau, q, t) \right) \|a_N W_{N:1}(0) - W_*\|_X^2 , \tag{53b}$$

$$\mathcal{E}_{DLN}(t) \leq 3\beta (q + \tau)^t \|a_N W_{N:1}(0) - W_*\|_X^2 , \tag{53c}$$

where $D(\tau, q, t) = \min \left\{ \frac{\tau}{1-q}, 2(q+\tau)^t \right\}$*, with* q *defined in (15).*

Proof of Lemma D.5. Using Lemma D.3, we obtain that for any $\tau \in (0, 1)$ and $\eta < \eta_1$, we can find small enough positive constant c_1, c_2, which are only depend on $\tau, \eta/\eta_1$, and constant $C = C(c_1, c_2) = C''(\tau, \eta/\eta_1)$ mentioned in Lemma D.3, such that

$$\eta = \frac{(1-\varepsilon) 2 n_N}{e^{6c_1 + 3c_2} \beta N},$$

where $0 < \varepsilon < 1$, as well as

$$V(t+1) = V(t) - \eta_* \nabla L(V(t)) + R(t),$$

where $V(t) = a_N W_{N:1}(t)$, $\eta_* = \frac{N}{n_N} \eta$, and $\|R(t)\|_X \leq \tau' = \tau \|V(t) - W_*\|_X$.

Notice that $\theta_0 := \eta/\eta_1 = \frac{1-\varepsilon}{e^{6c_1+3c_2}}$ and $\eta/\eta_0 = 1 - \varepsilon$, where $\eta_0 = \frac{2n_N}{e^{6c_1+3c_2}\beta N}$.

For the right hand side of inequality (32), we have

$$\frac{C(c_1, c_2) M^2 \kappa^2 B_0}{\varepsilon^2} N^{2\theta} n_N = \frac{C''(\tau, \eta/\eta_1) M^2 \kappa^2 B_0}{\varepsilon^2} N^{2\theta} n_N.$$

To show that inequality (32) is equivalent to inequality (52), it suffices to show that ε only depend on $\tau, \eta/\eta_1$. Notice that

$$\varepsilon = 1 - \eta/\eta_0 = 1 - \theta_0 e^{6c_1 + 3c_2},$$

and c_1, c_2 only depend on τ and η/η_1, which implies ε only depend on $\tau, \eta/\eta_1$.

Now, we start to prove the three inequalities in (53).

Recall the GD (2) for $W(t)$. Define $\Delta(t) = V(t) - W(t) = a_N W_{N:1}(t) - W(t)$. Notice that

$$\Delta(t+1) = \Delta(t) - \eta_* (\nabla L(V(t)) - \nabla L(W(t))) + R(t),$$

and

$$\begin{aligned} &\|\Delta(t+1)\|_X^2 \\ \leq &\eta_*^2 \|\nabla L(V(t)) - \nabla L(W(t))\|_X^2 - 2\eta_* \langle \Delta(t), \nabla L(V(t)) - \nabla L(W(t)) \rangle_X \\ &+ \|\Delta(t)\|_X^2 + (2 \|\Delta(t)\|_X + 2\eta_* \|\nabla L(V(t)) - \nabla L(W(t))\|_X + \tau') \tau'. \end{aligned}$$

Let $l_t = 2 \|\Delta(t)\|_X + 2\eta_* \|\nabla L(V(t)) - \nabla L(W(t))\|_X + \tau'$.

Now, we aim to find an upper bound for l_t.

Applying lemma C.2 with the assumption $0 < \eta_* = \frac{N}{n_N} \eta < \frac{2}{\beta}$, we know that

$$l_t \leq (6 \|\Delta(t)\|_X + \tau') \leq 7(\|W(t) - W_*\|_X + \|V(t) - W_*\|_X). \tag{54}$$

Thus

$$l_t \tau' \leq 7\tau \|V(t) - W_*\|_X (\|V(t) - W_*\|_X + \|W(t) - W_*\|_X) =: U_t \tau.$$

By inequality (23),

$$\begin{aligned}
&\|\Delta(t+1)\|_X^2 \\
\leq &\|\Delta(t)\|_X^2 - 2\eta_*\langle \Delta(t), \nabla L(V(t)) - \nabla L(W(t))\rangle_X \\
&+ \eta_*^2 \|\nabla L(V(t)) - \nabla L(W(t))\|_X^2 + U_t\tau \\
= &\|\Delta(t)\|_X^2 - 2\eta_*\langle V(t) - W(t), \nabla L(V(t)) - \nabla L(W(t))\rangle_X \\
&+ \eta_*^2 \|\nabla L(V(t)) - \nabla L(W(t))\|_X^2 + U_t\tau \\
\leq &\|\Delta(t)\|_X^2 - 2\eta_*\frac{\alpha\beta}{\alpha+\beta}\|\Delta(t)\|_X^2 \\
&+ \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right)\|\nabla L(V(t)) - \nabla L(W(t))\|_X^2 + U_t\tau.
\end{aligned}$$

Case 1: $\frac{2}{\alpha+\beta} < \eta_* < \frac{2}{\beta}$.
In this case, we have

$$\begin{aligned}
&\|\Delta(t+1)\|_X^2 \\
\leq &\|\Delta(t)\|_X^2 - 2\eta_*\frac{\alpha\beta}{\alpha+\beta}\|\Delta(t)\|_X^2 + \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right)\|\nabla L(V(t)) - \nabla L(W(t))\|_X^2 + U_t\tau \\
\leq &\|\Delta(t)\|_X^2 - 2\eta_*\frac{\alpha\beta}{\alpha+\beta}\|\Delta(t)\|_X^2 + \left(\eta_*^2 - \frac{2\eta_*}{\alpha+\beta}\right)\beta^2\|\Delta(t)\|_X^2 + U_t\tau \\
\leq &(1 - \beta\eta_*(2-\eta_*\beta))\|\Delta(t)\|_X^2 + U_t\tau \\
= &: q\|\Delta(t)\|_X^2 + U_t\tau.
\end{aligned}$$

Case 2: $0 < \eta_* \leq \frac{2}{\alpha+\beta}$.
Similarly, we have

$$\|\Delta(t+1)\|_X^2 \leq (1 - \alpha\eta_*(2-\eta_*\alpha))\|\Delta(t)\|_X^2 + U_t\tau =: q\|\Delta(t)\|_X^2 + U_t\tau. \tag{55}$$

In both cases, we have $0 < q < 1$.

First of all, since $U_t \leq U_0$ and $\|\Delta(0)\|_X = 0$, we obtain that

$$\|\Delta(t)\|_X^2 \leq \frac{U_0\tau}{1-q} + q^t\left(\|\Delta(0)\|_X^2 - \frac{U_0\tau}{1-q}\right) \leq \frac{U_0\tau}{1-q} \leq \frac{14\tau}{1-q}\|V(0) - W_*\|_X^2.$$

Applying Lemma D.4 for $V(t)$ and $W(t)$, we obtain $\|V(t) - W_*\|_X^2 \leq (1+\varepsilon)^t q^t \|V(0) - W_*\|_X^2$ and $\|W(t) - W_*\|_X^2 \leq q^t\|W(0) - W_*\|_X^2$, respectively. Thus,

$$\begin{aligned}
&|L(W(t)) - L(a_N W_{N:1}(t))| \\
\leq &|\langle \nabla L(W(t)), \Delta(t)\rangle_X| + \frac{\beta}{2}\|\Delta(t)\|_X^2 \\
\leq &\beta\|W(t) - W_*\|_X \cdot \|\Delta(t)\|_X + \frac{\beta}{2}\|\Delta(t)\|_X^2 \\
\leq &\beta\left(q^{t/2}\sqrt{\frac{14\tau}{1-q}} + \frac{7\tau}{1-q}\right)\|V(0) - W_*\|_X^2.
\end{aligned}$$

Generally speaking, (55) implies

$$\|\Delta(t)\|_X^2 \leq \tau\sum_{j=0}^{t-1} q^{t-1-j}U_j.$$

We have

$$\begin{aligned}
\|\Delta(t)\|_X^2 &\leq 14\tau\sum_{j=0}^{t-1}(q+7\tau)^j q^{t-1-j}\|V(0) - W_*\|_X^2 \\
&\leq 2(q+7\tau)^t\left(1 - \left(\frac{q}{q+7\tau}\right)^t\right)\|V(0) - W_*\|_X^2
\end{aligned}$$

Thus, we have

$$\|a_N W_{N:1}(t) - W(t)\|_X^2 \leq \min\left\{\frac{14\tau}{1-q}, 2(q+7\tau)^t\right\} \|V(0) - W_*\|_X^2 ,$$

as well as

$$\begin{aligned}
&|L(W(t)) - L(a_N W_{N:1}(t))| \\
\leq &\beta \|W(t) - W_*\|_X \cdot \|\Delta(t)\|_X + \frac{\beta}{2} \|\Delta(t)\|_X^2 \\
\leq &\beta \left(\sqrt{\min\left\{\frac{14\tau}{1-q}, 2(q+7\tau)^t\right\}} \cdot q^{t/2} + \frac{1}{2}\min\left\{\frac{14\tau}{1-q}, 2(q+7\tau)^t\right\}\right) \|V(0) - W_*\|_X^2 .
\end{aligned}$$

By triangle inequality as well as $L(W(t)) - L(W_*) \leq \frac{\beta}{2} q^t \|V(0) - W_*\|_X^2$, we have

$$|L(a_N W_{N:1}(t)) - L(W_*)| \leq 3\beta(q+7\tau)^t \|V(0) - W_*\|_X^2 .$$

Without loss of generality, we replace all 14τ and 7τ by τ, which completes the proof.

□

E GAUSSIAN INITIALIZATION FALL INTO THE CONVERGENCE REGION

In this section, we first establish some spectral properties of the products of random Gaussian matrices. The spectral properties lead to the conclusion that overparameterization guarantees that the random initialization will fall into the convergence region with high probability.

Gaussian initialization:

Denote by $N(0,1)$ the standard Gaussian distribution, and χ_k^2 the chi square distribution with k degrees of freedom. Let $S^{d-1} = \{x \in \mathbb{R}^d; \|x\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^d$.

The scaling factor $a_N = \frac{1}{\sqrt{n_1 n_2 \cdots n_N}}$ ensures that the networks at initialization preserves the norm of every input in expectation.

Lemma E.1. *For any $x \in \mathbb{R}^{n_0}$, the Gaussian initialization satisfies*

$$\mathbb{E}\left[\|a_N W_{N:1}(0)x\|_2^2\right] = \|x\|_2^2 .$$

Proof of Lemma E.1. For random matrix $A \in \mathbb{R}^{n_i \times n_{i-1}}$ with i.i.d $N(0,1)$ entries and any vector $0 \neq v \in \mathbb{R}^{n_{i-1}}$, the distribution of $\frac{\|Av\|_2^2}{\|v\|_2^2}$ is $\chi_{n_i^2}$. We rewrite

$$\|W_{N:1}(0)x\|_2^2 / \|x\|_2^2 = Z_N Z_{N-1} \cdots Z_1,$$

where $Z_i = \|W_{i:1}(0)x\|^2 / \|W_{i-1:1}(0)x\|^2$.

Then we know that the distribution of random variable $Z_1 \sim \chi_{n_1}^2$, and conditional distribution of random variables $Z_i|(Z_1, \cdots, Z_{i-1}) \sim \chi_{n_i}^2 (1 < i \leq N)$. Thus, $Z_1, \cdots, Z_{n_i}$ are independent. By law of iterated expectations, we have

$$\mathbb{E}[\|W_{N:1}(0)x\|_2^2 / \|x\|_2^2] = \prod_{j=1}^{N} n_j.$$

□

Define $\Delta_1 = \sum_{j=1}^{N-1} 1/n_j$. Now, we introduce a new notation $\Omega\left(\frac{1}{\Delta_1}\right)$, which means that there exists $k > 0$, such that $\Omega\left(\frac{1}{\Delta_1}\right) \geq \frac{k}{\Delta_1}$.

Lemma E.2. *Consider real random matrix $A_j \in \mathbb{R}^{n_j \times n_{j-1}}, 1 \leq j \leq q$ with i.i.d $N(0,1)$ entries and any vector $0 \neq x \in \mathbb{R}^{n_1}$.*
Define $\Delta_1(q) = \sum_{j=1}^{q} \frac{1}{n_j}$ and $n_{min} = \min_{1 \leq j \leq q} n_j$. Then

$$\mathbb{P}(\|A_q A_{q-1} \cdots A_1 x\|_2^2 / \|x\|_2^2 > e^c n_1 \cdots n_q) \leq \exp\left\{-\frac{c^2}{8\Delta_1(q)}\right\} =: f_1(c), \forall c > 0. \tag{56}$$

When $0 < c \leq 3\ln 2$, $\Delta_1(q) \leq c/(12\ln 2)$, *we have*

$$\mathbb{P}(\|A_q A_{q-1}\cdots A_1 x\|_2^2 / \|x\|_2^2 < e^{-c} n_1 \cdots n_q) \leq \exp\left\{-\frac{c^2}{36\ln(2)\Delta_1(q)}\right\} =: f_2(c). \tag{57}$$

Hence, for any $x \in S^{n_0-1}$ *with probability at least* $1 - e^{-\Omega(\frac{1}{\Delta_1(q)})}$, *we have*

$$e^{-c_2/2}(n_1\cdots n_q)^{1/2} \leq \|A_q\cdots A_1 x\|_2 \leq e^{c_1/2}(n_1\cdots n_q)^{1/2},$$

when $0 < c_2 \leq 3\ln 2$, $\Delta_1(q) \leq c_2/(12\ln 2)$.

Proof of Lemma E.2. For random matrix $A_i \in \mathbb{R}^{n_i \times n_{i-1}}$ with i.i.d $N(0,1)$ entries and any vector $0 \neq v \in \mathbb{R}^{n_{i-1}}$, the random variable $\frac{\|A_i v\|_2^2}{\|v\|_2^2}$ is distributed as $\chi^2_{n_i}$. We rewrite

$$\|A_q\cdots A_1 x\|_2^2 / \|x\|_2^2 = Z_q Z_{q-1}\cdots Z_1,$$

where $Z_i = \|A_{i:1}x\|^2 / \|A_{i-1:1}x\|^2$. We have $Z_1 \sim \chi^2_{n_1}$, $Z_i|(Z_1,\cdots,Z_{i-1}) \sim \chi^2_{n_i} (1 < i \leq q)$. Recall the moments of $Z \sim \chi^2_m$:

$$\mathbb{E}[Z^\lambda] = \frac{2^\lambda \Gamma(\frac{m}{2}+\lambda)}{\Gamma(\frac{m}{2})}, \forall \lambda > -\frac{m}{2}.$$

Now, we aim to find the Chernoff type bound.

Case 1: We define ratio of Gamma function

$$R(x,\lambda) = \frac{\Gamma(x+\lambda)}{\Gamma(x)}, \lambda > 0, x > 0.$$

In Jameson (2013), we have

$$R(x,\lambda) \leq x(x+\lambda)^{\lambda-1} \leq (x+\lambda)^\lambda, \lambda > 0, x > 0. \tag{58}$$

Fixed $c > 0$, for any $\lambda > 0$ we have

$$\begin{aligned}
\mathbb{P}(Z_q\cdots Z_1 > e^c n_1\cdots n_q) &\leq \mathbb{P}((Z_q\cdots Z_1)^\lambda > e^{\lambda c}(n_1\cdots n_q)^\lambda) \\
&\leq e^{-\lambda c}(n_1\cdots n_q)^{-\lambda}\mathbb{E}[(Z_q\cdots Z_1)^\lambda] && \text{(Markov inequality)} \\
&= \exp\{-\lambda(c+\ln(n_1\cdots n_q))\}\prod_{j=1}^q 2^\lambda R(n_j/2,\lambda) && \text{(Law of total expectation)} \\
&\leq \exp\{-\lambda(c+\ln(n_1\cdots n_q)) + q\lambda\ln 2 + \sum_{j=1}^q \lambda\ln(\frac{n_j}{2}+\lambda)\} && \text{(Inequality (58))} \\
&= \exp\{-\lambda c + \lambda\sum_{j=1}^q \ln(1+\frac{2\lambda}{n_j})\} \\
&\leq \exp\{-\lambda c + 2\lambda^2\sum_{j=1}^q \frac{1}{n_j}\}.
\end{aligned}$$

Define constant $\Delta_1(q) = \sum_{j=1}^q \frac{1}{n_j}$. Set $\lambda = \frac{c}{4\Delta_1(q)}$, we obtain (56).

Case 2: Let $n_{min} = \min_{1\leq j\leq q} n_j$.

$$\begin{aligned}
\mathbb{P}(Z_q\cdots Z_1 < e^{-c} n_1\cdots n_q) &\leq \mathbb{P}((Z_q\cdots Z_1)^\lambda > e^{-\lambda c}(n_1\cdots n_q)^\lambda) \\
&\leq \exp\{\lambda(c-\ln(n_1\cdots n_q)) + q\lambda\ln 2 + \sum_{j=1}^q \ln R(\frac{n_j}{2},\lambda)\}.
\end{aligned}$$

Define

$$f(\lambda) = \lambda(c - \ln(n_1 \cdots n_q)) + q\lambda \ln 2 + \sum_{j=1}^{q} \ln R(\frac{n_j}{2}, \lambda), -\frac{n_{min}}{2} < \lambda \leq 0.$$

Notice that $f(0) = 0$. Define digamma function,

$$\psi(x) = \frac{d}{dx} \ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Qi et al. (2006) proved the following sharp inequality of digamma function,

$$\ln(x + \frac{1}{2}) - \frac{1}{x} < \psi(x) < \ln(x + e^{-\gamma}) - \frac{1}{x}, x > 0,$$

where γ is the Euler-Mascheroni constant, and $e^{-\gamma} \approx 0.561459$.
Thus,

$$f'(\lambda) = c + \sum_{j=1}^{q} \left[-\ln(\frac{n_j}{2}) + \psi(\frac{n_j}{2} + \lambda) \right] \geq c + \sum_{j=1}^{q} \ln(1 + \frac{\lambda + 1/2}{n_j/2}) - \sum_{j=1}^{q} \frac{1}{n_j/2 + \lambda}.$$

Since $\ln(1 + x)$ is concave, we have

$$\ln(1 + x) \geq 2\ln(2)x, x \in [-1/2, 0].$$

If $-\frac{n_{min}}{4} \leq \lambda \leq 0$, then

$$\begin{aligned}
f(\lambda) &= f(0) - \int_{\lambda}^{0} f'(x)dx \\
&\leq c\lambda + \int_{0}^{\lambda} \left[\sum_{j=1}^{q} \ln(1 + \frac{x + 1/2}{n_j/2}) - \sum_{j=1}^{q} \frac{1}{n_j/2 + x} \right] dx \\
&= c\lambda + \sum_{j=1}^{q} \left[\lambda \ln(1 + \frac{\lambda + 1/2}{n_j/2}) + (n_j/2 + 1/2) \ln(1 + \frac{\lambda}{n_j/2 + 1/2}) - \lambda - \ln(1 + \frac{\lambda}{n_j/2}) \right] \\
&\leq c\lambda + \sum_{j=1}^{q} (\lambda - 1) \ln(1 + \frac{\lambda}{n_j/2}) \\
&\leq c\lambda + 4\ln(2)\lambda(\lambda - 1)\Delta_1(q).
\end{aligned}$$

Assume $0 < c \leq 3\ln 2$. Let $A = 12\ln 2$, and $\lambda^* = -\frac{c}{A\Delta_1(q)}$. Since $n_{min}\Delta_1(q) \geq 1$, we have $\lambda^* \geq -n_{min}/4$.
Assume $\Delta_1(q) \leq c/(12\ln 2)$.
Thus

$$f(\lambda^*) \leq -\frac{c^2}{A\Delta_1(q)} + 4\ln 2 \frac{c^2}{A\Delta_1(q)} \left(\frac{\Delta_1(q)}{c} + \frac{1}{A} \right) \leq -\frac{c^2}{36\ln(2)\Delta_1(q)}. \tag{59}$$

Thus, we obtain (57). □

Lemma E.3. *There exists a positive constant $C(c_1, c_2)$ which only depends on c_1, c_2, such that if $n_N\Delta_1 \leq C(c_1, c_2)$, then for any fixed $1 < i \leq N$, with probability at least $1 - \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\}$ we have*

$$\sigma_{\max}(W_{N:i}(0)) \leq e^{c_1}(n_{i-1}n_i \cdots n_{N-1})^{1/2}, \tag{60}$$

and

$$\sigma_{\min}(W_{N:i}(0)) \geq e^{-c_2}(n_{i-1}n_i \cdots n_{N-1})^{1/2}. \tag{61}$$

Proof of Lemma E.3. Let $A = W_{N:i}^T(0)$. We know that

$$\sigma_{max}(A) = \|A\| = \sup_{v \in S^{n_N - 1}} \|Av\|_2$$

and

$$\sigma_{min}(A) = \inf_{v \in S^{n_N - 1}} \|Av\|_2 .$$

Applying lemma E.2, we know that with probability at least $1 - \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\}$,

$$\|Av\|_2 / \|v\|_2 \in [e^{-c_2/2}P, e^{c_1/2}P],$$

where $P = (n_{i-1} \cdots n_{N-1})^{1/2}$.
Set $\phi = \min\{1 - e^{-c_1/2}, (e^{-c_2/2} - e^{-c_2})/(e^{-c_2/2} + e^{c_1})\}$. Take a ϕ-net $\mathcal{N}_\phi$ for $S^{n_N - 1}$ with size $|\mathcal{N}_\phi| \leq (3/\phi)^{n_N}$. Notice that with this size we can actually cover the unit ball, not only the unit sphere.
Thus, with probability at least $1 - |\mathcal{N}_\phi| \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\}$, for all $u \in \mathcal{N}_\phi$ simultaneously we have

$$\|Au\|_2 / \|u\|_2 \in [e^{-c_2/2}P, e^{c_1/2}P].$$

Fixed $v \in S^{n_N - 1}$, there exists $u \in \mathcal{N}_\phi$ such that $\|u - v\|_2 \leq \phi$. WLOG, we assume $1 - \phi \leq \|u\|_2 \leq 1$. We obtain

$$\|Av\|_2 \leq \|Au\|_2 + \|A(u - v)\|_2 \leq e^{c_1/2}P + \phi \|A\| .$$

Taking supereme over $\|v\|_2 = 1$, we obtain

$$\sigma_{max}(A) = \|A\| \leq \frac{e^{c_1/2}}{1 - \phi} P \leq e^{c_1} P.$$

For the lower bound, we have

$$\|Av\|_2 \geq \|Au\|_2 - \|A(u - v)\|_2 \geq e^{-c_2/2}P \|u\| - \phi \|A\| \geq \left[(1 - \phi)e^{-c_2/2} - \phi e^{c_1}\right] P \geq e^{-c_2} P.$$

Taking the infimum over $\|v\|_2 = 1$, we get

$$\sigma_{min}(A) \geq e^{-c_2} P.$$

The conclusions hold with probability at least

$$\begin{aligned} & 1 - |\mathcal{N}_\phi| \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\} \\ \geq & 1 - \exp\{n_N \ln(3/\phi)\} \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\} \\ \geq & 1 - \exp\left\{-\Omega\left(\frac{1}{\Delta_1}\right)\right\}, \end{aligned}$$

since $n_N \Delta_1 \leq C(c_1, c_2)$. □

Lemma E.4. *There exists a positive constant $C(c_1, c_2)$ which only depends on c_1, c_2, such that if $rank(X)\Delta_1 \leq C(c_1, c_2)$, then for any fixed $1 \leq j < N$, with probability at least $1 - \exp\{-\Omega\left(\frac{1}{\Delta_1}\right)\}$ we have*

$$\sigma_{\max}(W_{j:1}(0)|_{\mathcal{R}(X)}) \leq e^{c_1}(n_1 n_2 \cdots n_j)^{1/2}, \tag{62}$$

and

$$\sigma_{\min}(W_{j:1}(0)|_{\mathcal{R}(X)}) \geq e^{-c_2}(n_1 n_2 \cdots n_j)^{1/2}. \tag{63}$$

Proof of Lemma E.4. The proof is similar to that of previous lemma. The only difference is that now we consider the $\phi-$net to cover the unit sphere in $\mathcal{R}(X) \cap \mathbb{R}^{n_0}$, with $\dim \mathcal{R}(X) \cap \mathbb{R}^{n_0} = rank(X)$, where $\mathcal{R}(X)$ represents the column space of X. □

Lemma E.5. *Set $C = n_{max}/n_{min} < \infty$, $\theta = 1/2$. Assume $\Omega(1/\Delta_1) \geq \frac{k}{\Delta_1}$, where $0 < k < 1$ is a constant and Δ_1 satisfies*

$$\begin{cases} \Delta_1 \leq \min\left\{\frac{k}{5\ln(6)}, \frac{k}{5\ln(5\ln(6)e/k)}\right\} \\ \Delta_1 \ln(C) \leq \min\left\{\frac{k}{5\ln(5\ln(6)e/k)}, \frac{k}{5}\right\} \\ \Delta_1 \ln(N^{2\theta}) \leq k/5. \end{cases}$$

Given $1 < i \leq j < N$, with probability at least $1 - 2e^{-k/(5\Delta_1)} = 1 - e^{-\Omega(1/\Delta_1)}$ we have

$$\|W_{j:i}(0)\| \leq M_k \sqrt{C} N^{\theta} (n_i \cdots n_{j-1} \cdot \max\{n_{i-1}, n_j\})^{1/2},$$

where M_k is a positive constant that only depends on k.

Proof of Lemma E.5. WLOG, assume $n_{i-1} \leq n_j$. Let $A = W_{j:i}(0)$. From lemma E.2, we know that fixed $v \in S^{n_{i-1}-1}$, with probability at least $1 - e^{-\Omega(1/\Delta_1)}$ we have $\|Av\|_2 \leq 4/3(n_i \cdots n_j)^{1/2}$.
.
Take a small constant $c = \frac{kN^{2\theta}}{5\ln(6)\Delta_1 n_{i-1}} \geq \frac{k}{5\ln(6)C}$. Let $v_1, \cdots, v_{n_{i-1}}$ be an orthonormal basis for $R^{n_{i-1}}$. Partition the index set $\{1, 2, \cdots, v_{n_{i-1}}\} = S_1 \cup S_2 \cup \cdots \cup S_{\lceil N^{2\theta}/c \rceil}$, where $|S_l| \leq \lceil cn_{i-1}/N^{2\theta} \rceil$ for each $1 \leq l \leq \lceil N^{2\theta}/c \rceil$.
The following discussion is similar to the proof of lemma E.3, hence we omit some details. For each l, taking a $1/2-$ net $\mathcal{N}_l$ for the set $V_{S_l} = \{v \in S^{n_{i-1}-1}; v \in span\{v_i; i \in S_l\}\}$, we can get

$$\|Au\|_2 \leq 4(n_i \cdots n_j)^{1/2}, u \in V_{S_l},$$

with probability at least

$$1 - |\mathcal{N}_l| e^{-k/\Delta_1} \geq 1 - \exp\{-k/\Delta_1 + (cn_{i-1}/N + 1)\ln 6\} \geq 1 - e^{-3k/(5\Delta_1)},$$

since $\Delta_1 \leq \frac{k}{5\ln(6)}$.
Therefore, for any $v \in \mathbb{R}^{n_{i-1}}$, we can write it as the sum $v = \sum_l a_l v_l$, where $\alpha_l \in \mathbb{R}$ and $v_l \in V_{S_l}$ for each l. We also know that $\|v\|_2^2 = \sum_{l \geq 1} |\alpha_l|^2$.
Then we have

$$\|Av\|_2 \leq \sum_l |\alpha_l| \|Av_l\|_2 \leq 4(n_i \cdots n_j)^{1/2} \sqrt{\lceil N^{2\theta}/c \rceil \sum_l |a_l|^2} \leq M_k \sqrt{C} N^{\theta} (n_i \cdots n_j)^{1/2} \|v\|_2 .$$

Thus,

$$\|A\| \leq M_k \sqrt{C} N^{\theta} (n_i \cdots n_j)^{1/2}.$$

Notice that when $C \leq e$, $\Delta_1 \leq \frac{k}{5\ln(5\ln(6)e/k)} \leq \frac{k}{5\ln(5\ln(6)\cdot C/k)}$, and when $C > e$, we have

$$\Delta_1 \ln(C) \leq \min\left\{\frac{k}{5\ln(5\ln(6)e/k)}, k/5\right\} \leq \frac{k\ln(C)}{5\ln(5\ln(6)\cdot C/k)}.$$

The success probability is at least

$$\begin{aligned} & 1 - \lceil N^{2\theta}/c \rceil \cdot e^{-3k/(5\Delta_1)} \\ \geq & 1 - \exp\left\{\ln\left(\frac{5\ln(6)\cdot C}{k}\right) + \ln(N^{2\theta}) - 3k/(5\Delta_1)\right\} - e^{-3k/(5\Delta_1)} \\ \geq & 1 - 2e^{-k/(5\Delta_1)}, \end{aligned}$$

since

$$\Delta_1 \leq \frac{k}{5\ln(5\ln(6)\cdot C/k)} \text{ and } \Delta_1 \ln(N^{2\theta}) \leq k/5.$$

□

Proof of Lemma 2.3. Set $r = rank(X)$, and $u_1, \cdots, u_r$ be an orthonormal basis of the column space of X.

Then, $P_X = \sum_{i=1}^{r} u_i u_i^T$.

Notice that

$$\|a_n W_{N:1}(0)\|_X^2 = \|a_n W_{N:1}(0) P_X\|_F^2 = \sum_{i=1}^{r} \|a_n W_{N:1}(0) u_i\|_2^2 .$$

By assumption, we have

$$\mathbb{E}\,\|a_n W_{N:1}(0)\|_X^2 = \mathbb{E} \sum_{i=1}^{r} \|a_n W_{N:1}(0) u_i\|_2^2 = r.$$

The Markov inequality implies

$$\mathbb{P}(\|a_n W_{N:1}(0)\|_X^2 \geq \frac{2r}{\delta}) \leq \frac{\delta}{2}.$$

Therefore, we can bound the initial loss value as

$$\begin{aligned} L_0 - L(W_*) &\leq \langle \nabla L(W_*), a_N W_{N:1}(0) X - W_* \rangle + \frac{\beta}{2} \|a_N W_{N:1}(0) - W_*\|_X^2 \\ &= \frac{\beta}{2} \|a_N W_{N:1}(0) - W_*\|_X^2 \\ &\leq \beta(\|a_N W_{N:1}(0)\|_X^2 + \|W_*\|_X^2) \\ &\leq \beta(\frac{2r}{\delta} + \|W_*\|_X^2), \end{aligned}$$

with probability at least $1 - \delta/2$.

□

Proof of Theorem B.1. The requirement on size $\{n_1, n_2, \cdots, n_{N-1}, N\}$ in (16) makes sure that lemma E.3, E.4, E.5, 2.3, and D.1 hold.

WLOG, we set $c_1 = c/6, c_2 = c/3$, $M = 2M_k\sqrt{C_0}$, $B_0 = B_\delta$, and $\eta =: \frac{(1-\varepsilon)2n_N}{e^{2c}\beta N}$, then with probability at least

$$1 - N^2 e^{-\Omega(1/\Delta_1)} - \delta/2 \geq 1 - \delta, \text{ since } \Delta_1 \leq \frac{1}{C(c)} \min\left\{\frac{1}{\ln N}, \frac{1}{\ln(1/\delta)}\right\},$$

the random initialization satisfies the initialization assumption (31) and the overparameterization assumption (32). Applying Lemma D.1, we complete the proof.

□

F ORTHOGONAL INITIALIZATION FALL INTO THE CONVERGENCE REGION

There are some basic facts for random projections and embeddings. Most of the following properties can be found in Eaton (1989).

Proposition F.1.

1. *A is a random embedding if and only if A^T is a random projection.*
2. *If A is a square matrix, then random projection, random embedding and random orthogonal matrix are equivalent.*
3. *The uniform distribution on the group is a left and right invariant probability measure, that is, if A is a random orthogonal matrix, then A, UA, AU are all random orthogonal matrix, where U is a non-random orthogonal matrix.*

4. *Assume X is a $n \times q (q \leq n)$ random matrix whose entries are i.i.d. $N(0,1)$ random variables. Then $A := X(X^TX)^{-1/2}$ is a random embedding, since $A^TA = I_q$ and the distribution of A is left invariant, which means that A and UA have the same distribution, where U is a non-random orthogonal matrix.*

5. *If A is a uniform distribution over an orthogonal group of order n and A is partitioned as $A = (A_1, A_2)$, where A_1 is $n \times q$ and A_2 is $n \times (n-q)$, then A_1^T and A_2^T are both random orthogonal matrix.*

6. *The columns of uniform distribution over orthogonal group of order n, and*

$$\frac{(\xi_1, \cdots, \xi_n)}{\sqrt{\xi_1^2 + \xi_2^2 + \cdots + \xi_n^2}}$$

have the same distribution, where $\xi_1, \cdots, \xi_n$ are i.i.d. $N(0,1)$ random variables.

7. *Assume $A = A_{n\times p}, n \leq p$ is a random orthogonal projection. For any $v \in S^{p-1}$, $\|Av\|_2^2$ and $(\sum_{i=1}^n \xi_i^2)/(\sum_{j=1}^p \xi_j^2)$ are both following beta distribution with $\alpha = n/2, \beta = (p-n)/2$, where $\xi_1, \cdots, \xi_n$ are i.i.d. $N(0,1)$ random variables.*

Remark 9. There are several ways to construct random matrix $A = (a_{ij})_{q\times n},\ q \leq n$, which is uniformly distributed over rectangular matrices with $AA^T = c^2 I_q, c > 0$. Let O_n be uniformly distributed over real orthogonal group of order n, and O_n is partitioned as $O_n = (A_1^T, A_2^T)^T$, where A_1 is $q \times n$. Assume $X = (x_{ij})_{q\times n}$, and x_{ij} are independent standard normal random variables. Then A, cA_1, and $c(XX^T)^{-1/2}X$ have the same distribution.

Lemma F.2. *For any $x \in \mathbb{R}^{n_0}$, the one peak random projections and embedding initiation satisfies*

$$\mathbb{E}\left[\|a_N W_{N:1}(0)x\|_2^2\right] = \|x\|_2^2 .$$

Proof. Let $D = W_{p:1}(0)/\sqrt{n_1 n_2 \cdots n_p}$. Then D is an embedding matrix. Thus, $\|Dx\|_2^2 = \|x\|_2^2$. Let $A_i = W_{i:p+1}(0)/\sqrt{n_p n_{p+1} \cdots n_{i-1}}$, where $i \geq p+1$, and $A_p = I$.

Set $B_i = \|A_i Dx\|_2^2 / \|A_{i-1}Dx\|_2^2$, $i \geq p+1$. Then, B_i follows beta distribution $B(n_i/2, (n_{i-1} - n_i)/2)$ given $B_{i-1}, B_{i-2}, \cdots, B_{p+1}$, $i \geq p+1$. If $n_i = n_{i-1}$, then $B_i | (B_{i-1}, B_{i-2}, \cdots, B_{p+1}) = 1$, a.s.

If $B \sim B(a,b)$, then the expectation is given by the following equation,

$$\mathbb{E}B = \frac{a}{a+b}.$$

Thus, by law of total expectation, we have

$$\frac{n_N}{n_p}\mathbb{E}\|a_N W_{N:1}(0)x\|_2^2 = \mathbb{E}\|A_N Dx\|_2^2 = \mathbb{E}B_N B_{N-1}\cdots B_{p+1}\|Dx\|_2^2 = \frac{n_N}{n_p}\|x\|_2^2 .$$

This completes the proof. □

Next, we introduce sub-Gaussian random variables, associated with bounds on how a random variables deviate their expected value.

Definition F.1. A random variable X with finite mean $\mu = \mathbb{E}X$ is sub-Gaussian if there is a positive number σ such that:

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \text{ for all } \lambda \in \mathbb{R} \tag{64}$$

Such a constant σ^2 is called a proxy variance, and we say that X is σ^2-sub-Gaussian, and we write $X \sim SG(\sigma^2)$.

Example F.1. Normal distribution $N(\mu, \sigma^2)$ of course is σ^2 sub-Gaussian.
For beta distribution, Elder (2016) showed that $B(a,b)$ is $\frac{1}{4(a+b)+2}$-sub-Gaussian and later, Marchal & Arbel (2017) concluded $\frac{1}{4(a+b+1)}$-sub-Gaussian.

The Hoeffding bound for random variable X with mean μ and sub-Gaussian parameter σ is given by,

$$\mathbb{P}\left[|X-\mu| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{2\sigma^2}\right\}, \forall t \geq 0. \tag{65}$$

Simply applying the Chernoff bound for $B(a,b)$, we obtain the following lemma.

Lemma F.3. *Assume random variable B distributed as beta distribution $B(a,b)$ with two positive shape parameters a and b. Then*

$$\mathbb{P}(\left|B-\frac{a}{a+b}\right| \geq y) \leq 2\exp\left\{-2(a+b)y^2\right\}, y \geq 0.$$

Hence,

$$\mathbb{P}\left(\left|B-\frac{a}{a+b}\right| \leq \varepsilon\frac{a}{a+b}\right) \geq 1-\exp\{-\Omega(a^2/(a+b))\},$$

where $\Omega(\cdot)$ only depend on ε.
For the upper tail, we can obtain a better bound,

$$\mathbb{P}\left(B \geq (1+\varepsilon)\frac{a}{a+b}\right) \leq \exp\left\{-(\varepsilon-\ln(\varepsilon+1))a\right\}. \tag{66}$$

Proof of Lemma F.3. We only need to prove the third inequality. Assume random variable $B \sim B(a,b)$. Set $v=a+b$, $(1+t)\frac{a}{v} \leq y < 1, t > 0$, and $r > 0$.
We are going to estimate the Chernoff bound for B, which is

$$\mathbb{P}(B \geq y) \leq e^{-(ry-\ln \mathbb{E}e^{rB})} =: e^{-I_r(y)}.$$

The moment generating function of B is given by

$$\mathbb{E}e^{rB} = 1+\sum_{k=1}^{\infty}\frac{a(a+1)\cdots(a+k-1)}{v(v+1)\cdots(v+k-1)}\frac{r^k}{k!} \leq 1+\sum_{k=1}^{\infty}\frac{a(a+1)\cdots(a+k-1)}{v^k}\frac{r^k}{k!}, r > 0.$$

Recall that the Maclaurin series of $(1-r/v)^{-a}$ over $(-v,v)$, is given by equation

$$(1-r/v)^{-a} = 1+\sum_{k=1}^{\infty}\frac{a(a+1)\cdots(a+k-1)}{v^k}\frac{r^k}{k!}.$$

Thus,

$$I_r(y) = ry-\ln \mathbb{E}e^{rB} \geq ry+a\ln(1-r/v).$$

Set $r = v-a/y \in (0,v)$. We obtain

$$\mathbb{P}(B \geq y) \leq \exp\{-(vy-a+a\ln(a/(vy)))\} =: \exp\{-vy\cdot g(a/(vy))\}, (1+t)\frac{a}{v} \leq y < 1$$

where $g(x) = 1-x+x\ln(x)$, $x = a/(vy) \in (0, 1/(1+t)]$. Notice that $g(1) = 0$ and $g'(x) = \ln(x) < 0$ over $x \in (0,1)$.
We know that

$$g(x) \geq g(1/(1+t)) = \frac{t-\ln(1+t)}{t+1}, t > 0.$$

Thus,

$$\mathbb{P}(B \geq y) \leq \exp\left\{-vy\cdot\frac{t-\ln(1+t)}{t+1}\right\} = \exp\left\{-(t-\ln(1+t))a\right\}, y = (1+t)\frac{a}{v} < 1.$$

Set $y = (1+\varepsilon)\frac{a}{a+b}$. We obtain the inequality (66). □

Remark 10. It is trivial to check

$$\|W_{j:i}(0)\| = (n_i n_{i+1}\cdots n_j)^{1/2}, 1 \leq i \leq j \leq p,$$
$$\|W_{j:i}(0)\| = (n_{i-1} n_i \cdots n_{j-1})^{1/2}, p+1 \leq i \leq j \leq N,$$
$$\|W_{j:i}(0)\| \leq (n_i n_{i+1}\cdots n_{j-1})^{1/2}(n_p)^{1/2}$$
$$\leq \left(\frac{n_{max}}{n_{min}}\right)^{1/2}(n_i n_{i+1}\cdots n_{j-1}\cdot\max\{n_{i-1},n_j\})^{1/2}, 1 \leq i < p < j \leq N, (i,j) \neq (1,N).$$

Remark 11. As a special case, if $n_1 = n_2 = \cdots = n_{N-1} = n$, we know that $\|W_{j:i}(0)\| = (n_{i-1}n_i \cdots n_{N-1})^{1/2} = n^{(N-i+1)/2}$.

Lemma F.4. *Assume $n_p/\min\{n_1, n_{N-1}\} \leq C_0 < \infty$. Set $\varepsilon > 0$. Let $C(\varepsilon)$ represent the constant depend only on ε. If $n_1/C_0 \geq C(\varepsilon)n_N$, then with probability at least $1 - e^{-\Omega(n_1/C_0)}$*

$$\sigma_{max}(W_{N:i}(0)) \leq (1+\varepsilon)(n_{i-1}n_i \cdots n_{N-1})^{1/2}, 2 \leq i \leq p$$
$$\sigma_{min}(W_{N:i}(0)) \geq (1-\varepsilon)(n_{i-1}n_i \cdots n_{N-1})^{1/2}, 2 \leq i \leq p.$$

Similarly, if $n_{N-1}/C_0 \geq C(\varepsilon)rank(X)$, then with probability at least $1 - e^{-\Omega(n_{N-1}/C_0)}$

$$\sigma_{max}(W_{j:1}(0)|_{\mathcal{R}(X)}) \leq (1+\varepsilon)(n_1 n_2 \cdots n_j)^{1/2}, p+1 \leq j \leq N$$
$$\sigma_{min}(W_{j:1}(0)|_{\mathcal{R}(X)}) \geq (1-\varepsilon)(n_1 n_2 \cdots n_j)^{1/2}, p+1 \leq j \leq N.$$

Proof of Lemma F.4. Let $D = (n_{N-1}n_{N-2}\cdots n_p)^{-1/2}W_{N:p+1}^T(0)$ and $A_i = (n_p n_{p-1}\cdots n_i)^{-1/2}W_{p:i}^T(0)$. Assume $v \in S^{n_N-1}$. Easy to see that A_i is a product of random orthogonal projections and D is a random embedding.
Let $e_1 = (1, 0, 0, \cdots, 0)^T \in \mathbb{R}^{n_p}$. There exists orthogonal matrix T such that $TDv = e_1$, $\|e_1\|_2 = \|TDv\|_2 = \|v\|_2 = 1$.
Since random orthogonal projections are right invariant, we have

$$\mathbb{P}(\|A_i Dv\|_2 \geq y) = \mathbb{E}\left[\mathbb{E}\left(I_{\{\|A_i T^T e_1\|_2 \geq y\}}\middle| D\right)\right] = \mathbb{E}\left[\mathbb{E}\left(I_{\{\|A_i e_1\|_2 \geq y\}}\middle| D\right)\right] = \mathbb{P}(\|A_i e_1\|_2 \geq y).$$

This proves that $\|A_i Dv\|_2^2$ and $\|A_i e_1\|_2^2$ have the same distribution.

Claim: If $v \neq 0$, then $\|A_i Dv\|_2^2 / \|v\|_2^2 = \left\|(n_i n_{i+1}\cdots n_p^2 \cdots n_{N-1})^{-1/2}W_{N:i}^T v\right\|_2^2 / \|v\|_2^2$ follows beta distribution $B(n_{i-1}/2, (n_p - n_{i-1})/2)$.
Define $B_p = \|A_p e_1\|_2^2$, $B_i = \|A_i e_1\|_2^2 / \|A_{i+1}e_1\|_2^2$, $i = p-1, p-2, \cdots, 1$.
Then $B_p \sim B(n_{p-1}/2, (n_p - n_{p-1})/2)$, $B_{p-1}|B_p \sim B(n_{p-2}/2, (n_{p-1} - n_{p-2})/2)$, $\cdots$, $B_i|(B_p, \cdots, B_{i+1}) \sim B(n_{i-1}/2, (n_i - n_{i-1})/2)$.
If $n_{i+1} = n_i$, we know that $B_i|(B_p, \cdots, B_{i+1}) = 1, a.s.$
If $B \sim B(a, b)$, then the moments are given by the following equations,

$$\mathbb{E}B = \frac{a}{a+b}, \text{ and } \mathbb{E}B^k = \frac{a}{a+b}\frac{a+1}{a+b+1}\cdots\frac{a+k-1}{a+b+k-1}. \tag{67}$$

By law of total expectation, we have

$$\mathbb{E}B_i B_{i+1}\cdots B_p = \frac{n_{i-1}}{n_i}\frac{n_i}{n_{i+1}}\cdots\frac{n_{p-1}}{n_p} = \frac{n_{i-1}}{n_p},$$

as well as

$$\mathbb{E}(B_i B_{i+1}\cdots B_p)^k = \frac{n_{i-1}/2}{n_p/2}\frac{n_{i-1}/2+1}{n_p/2+1}\cdots\frac{n_{i-1}/2+k-1}{n_p/2+k-1}.$$

Notice that all integer moments of $B_i B_{i+1}\cdots B_p$ match those of $B(n_{i-1}/2, (n_p - n_{i-1})/2)$. We can verify that beta distribution satisfies Carleman's condition, which implies that $B_i B_{i+1}\cdots B_p \sim B(n_{i-1}/2, (n_p - n_{i-1})/2)$.
Thus, $\|A_i Dv\|_2^2 / \|v\|_2^2 \sim B(n_{i-1}/2, (n_p - n_{i-1})/2)$, which proves the claim.

With probability at least $1 - \exp\{-\Omega(n_1/C_0)\}$, we have

$$(1-\varepsilon)^2\frac{n_{i-1}}{n_p} \leq \|ADv\|_2^2 \leq (1+\varepsilon)^2\frac{n_{i-1}}{n_p}, \|v\|_2 = 1.$$

Using the $\phi-$net technique which has already been used to prove lemma E.3, we know that

$$\sigma_{min}(AD) \geq (1-\varepsilon)\left(\frac{n_{i-1}}{n_p}\right)^{1/2},$$

and

$$\sigma_{max}(AD) \leq (1+\varepsilon)\left(\frac{n_{i-1}}{n_p}\right)^{1/2},$$

with probability at least $1-\exp\{n_N \ln(3/\phi(\varepsilon))\}\exp\{-\Omega(n_1/C_0)\} \geq 1-\exp\{-\Omega(n_1/C_0)$, since $n_1/C_0 \geq C(\varepsilon)n_N$, for $2 \leq i \leq p$.
Hence, with probability at least $1-e^{-\Omega(n_1/C_0)}$, we have

$$\sigma_{min}(W_{N:i}(0)) \geq (1-\varepsilon)\left(n_{i-1}\cdots n_{N-1}\right)^{1/2},$$

and

$$\sigma_{max}(W_{N:i}(0)) \leq (1+\varepsilon)\left(n_{i-1}\cdots n_{N-1}\right)^{1/2}.$$

The other part of the proof is similar to that of lemma E.4, so we omit it.

□

Proof of Theorem B.2 . Set $c>0$, $c_1 = c/6, c_2 = c/3$. In lemma F.4, we can pick a $\varepsilon > 0$, such that $1+\varepsilon \leq e^{c_1/2}$ and $1-\varepsilon \geq e^{-c_2/2}$. Set $M = 2\sqrt{C_0}, \theta = 0$, $B_0 = B_\delta$, and $\eta = \frac{(1-\varepsilon)2n_N}{e^{2c}\beta N}$.
The requirement on size $\{n_1, n_2, \cdots, n_{N-1}, N\}$ in (17) make sure that the remark 10, lemma F.4, lemma 2.3, and lemma D.1 all hold.
Notice that even though we need the conclusions in lemma F.4 simultaneously hold for $2 \leq i \leq p$, $p+1 \leq j \leq N$, it suffices to apply lemma F.4 over $i \in I$ and $j \in J$, such that $\{n_i; i \in I\}$ and $\{n_j; j \in I\}$ both have distinct values. Since $|I| \leq \underline{N}$ and $|J| \leq \underline{N}$, with probability at least

$$1 - 2\underline{N}e^{-\Omega(n_{min}/C_0)} - \delta/2 \geq 1-\delta,$$

the one peak random orthogonal projections and embeddings initialization satisfies the initialization assumption (31) and the overparameterization assumption (32).
Under assumption $n_1 = n_2 = \cdots = n_{N-1}$, we can use remark 11 to replace lemma F.4. Thus, with probability at least $1-\delta/2 \geq 1-\delta$, (31) holds. Applying lemma 2.3 and D.1, we complete the proof.

□

Proof of Theorem B.3. Let $W_N(0) = \sqrt{n}U_N[I_{n_y}, 0]V_N^T, \cdots, W_i(0) = \sqrt{n}U_i I_n V_i^T, 2 \leq i \leq N-1$, and $W_1(0) = \sqrt{n}U_1[I_{n_x}, 0]^T V_1^T$. Now, we want to verify (31). By simply calculation, we have

$$\begin{cases} \sigma_{max}(W_{N:i+1}(0)) = \sigma_{min}(W_{N:i+1}(0)) = n^{(N-i)/2}, 1 \leq i \leq N-1, \\ \sigma_{max}(W_{i-1:1}(0)|_{\mathcal{R}(X)}) = \sigma_{max}(W_{i-1:1}(0)|_{\mathcal{R}(X)}) = n^{(i-1)/2}, 2 \leq i \leq N, \\ \|W_{j:i}(0)\| = n^{(j-i+1)/2}, 1 < i \leq j < N. \end{cases} \tag{68}$$

Notice that for any $1 \leq p \leq m$

$$\|a_N W_{N:1}(0)x\|_2^2 = \frac{n}{n_N}\left\|U_N[I_{n_y},0]V_N^T U_N[I_{n_x},0]^T V_1^T x\right\|_2^2 = \frac{n}{n_N}\left\|U_N[I_{n_y},0]V_N^T x'\right\|_2^2,$$

where $x' = U_N[I_{n_x},0]^T V_1^T x$, $\|x\|_2 = \|x'\|_2$.
Since the distribution of $U_N[I_{n_y},0]V_N^T$ is right invariant under multiplying orthogonal matrices, we have

$$\left\|U_N[I_{n_y},0]V_N^T x'\right\|_2^2 / \|x\|_2^2 \sim B(\frac{n_y}{2}, \frac{n-n_y}{2}).$$

Thus,

$$\mathbb{E}\left[\|a_N W_{N:1}(0)x\|_2^2\right] = \|x\|_2^2.$$

Applying lemma 2.3, we have

$$L_0 - L(W_*) \leq \beta\left(\frac{2\cdot rank(X)}{\delta} + \|W_*\|_X^2\right),$$

with probability at least $1-\delta/2$.
Applying Lemma D.1 with $c>0$, $c_1 = c/6, c_2 = c/3$, $\theta = 0$, we complete the proof. □

Proof of Theorem 3.1. Theorem 3.1 is a special case of Theorem B.1 and Theorem B.2. Hence, we omit the proof. □

Proof of Theorem 3.2. In Theorem B.1, B.2, and B.3, we proved that for given constant $c_1, c_2 > 0$ and $0 < \varepsilon, \delta/2 < 1/2$ as well as learning rate η, there exists constant $C = C(c_1, c_2)$ such that all three kinds of random initializations will fall into the convergence region with probability at least $1 - \delta$. Applying Lemma 2.3, we complete the proof.

□

G TABLES

In this section, we provide some empirical evidence to support the argument in Section 4:**Why do bad saddles not affect GD for overparameterized deep linear neural networks?** Consider the following procedures for tables of $\frac{\|W_i(t)-W_i(0)\|_F}{\|W_i(0)\|_F}$:

a) We consider $X \in \mathbb{R}^{128\times 1000}$, and $W_* \in \mathbb{R}^{10\times 128}$ and set $Y = W_*X + \varepsilon$, where the entries in X and ε are drawn i.i.d. from $N(0, 1)$.

b) We consider the loss function $\frac{1}{2}\|a_N W_{N:1}X - Y\|_F^2$.

c) For the given deep linear networks, we apply orthogonal initialization, which are denoted as $W_j(0), 1 \le j \le N$.

d) We set the learning rate $\eta = \frac{n_N}{N\cdot\|X\|^2}$ for the deep linear neural networks.

e) We make the tables for $\frac{\|W_i(t)-W_i(0)\|_F}{\|W_i(0)\|_F}$.

Let $n_1 = n_2 = n_3 = 2000, N = 4$. Assume W_* are drawn i.i.d. from $N(0, 25)$. We obtain the following table:

	$i=1$	$i=2$	$i=3$	$i=4$
$t=1$	0.05161	0.00826	0.00826	0.18464
$t=2$	0.08779	0.01389	0.01389	0.31396
$t=3$	0.11335	0.01781	0.01779	0.40435
$t=4$	0.12109	0.01894	0.01889	0.42920
$t=5$	0.12527	0.01956	0.01948	0.44282
$t=6$	0.12611	0.01967	0.01958	0.44476
$t=7$	0.12755	0.01988	0.01978	0.44955
$t=8$	0.12745	0.01986	0.01975	0.44876
$t=9$	0.12819	0.01997	0.01987	0.45136
$t=10$	0.12793	0.01992	0.01982	0.45018

Let $n_1 = n_2 = 10000, N = 2$. Assume W_* are drawn i.i.d. from $N(0, 4)$. We obtain the following table:

	$i=1$	$i=2$	$i=3$
$t=1$	0.02708	0.00153	0.04844
$t=2$	0.04319	0.00244	0.07727
$t=3$	0.05296	0.00299	0.09474
$t=4$	0.05888	0.00333	0.10533
$t=5$	0.06248	0.00353	0.11176
$t=6$	0.06468	0.00365	0.11569
$t=7$	0.06603	0.00373	0.11811
$t=8$	0.06688	0.00377	0.11962
$t=9$	0.06741	0.00380	0.12057
$t=10$	0.06775	0.00382	0.12117

Let $n_1 = n_2 = 4000, N = 2$. Assume W_* are drawn i.i.d. from $N(0, 1)$. We obtain the following table:

	$i = 1$	$i = 2$	$i = 3$
$t = 1$	0.01622	0.00290	0.05802
$t = 2$	0.02684	0.00480	0.09601
$t = 3$	0.03411	0.00609	0.12202
$t = 4$	0.03919	0.00700	0.14018
$t = 5$	0.04280	0.00764	0.15306
$t = 6$	0.04539	0.00810	0.16232
$t = 7$	0.04729	0.00844	0.16908
$t = 8$	0.04869	0.00868	0.17408
$t = 9$	0.04974	0.00887	0.17782
$t = 10$	0.05054	0.00901	0.18066

Let $n_1 = n_2 = 8000, N = 2$. Assume W_* are drawn i.i.d. from $N(0, 1)$. We obtain the following table:

	$i = 1$	$i = 2$	$i = 3$
$t = 1$	0.01173	0.00148	0.04195
$t = 2$	0.01944	0.00246	0.06955
$t = 3$	0.02470	0.00312	0.08838
$t = 4$	0.02838	0.00358	0.10151
$t = 5$	0.03098	0.00391	0.11083
$t = 6$	0.03287	0.00415	0.11758
$t = 7$	0.03426	0.00432	0.12253
$t = 8$	0.03530	0.00445	0.12624
$t = 9$	0.03608	0.00455	0.12904
$t = 10$	0.03668	0.00463	0.13118

Let $n_1 = n_2 = 12000, N = 2$. Assume W_* are drawn i.i.d. from $N(0, 1)$. We obtain the following table:

	$i = 1$	$i = 2$	$i = 3$
$t = 1$	0.00965	0.00099	0.03453
$t = 2$	0.01597	0.00164	0.05712
$t = 3$	0.02025	0.00208	0.07244
$t = 4$	0.02323	0.00239	0.08310
$t = 5$	0.02535	0.00261	0.09069
$t = 6$	0.02690	0.00277	0.09621
$t = 7$	0.02804	0.00289	0.10029
$t = 8$	0.02890	0.00297	0.10336
$t = 9$	0.02955	0.00304	0.10570
$t = 10$	0.03006	0.00309	0.10750

Let $n_1 = n_2 = 20000, N = 2$. Assume W_* are drawn i.i.d. from $N(0, 1)$. We obtain the following table:

	$i = 1$	$i = 2$	$i = 3$
$t = 1$	0.00713	0.00057	0.02551
$t = 2$	0.01181	0.00095	0.04225
$t = 3$	0.01499	0.00121	0.05362
$t = 4$	0.01720	0.00138	0.06154
$t = 5$	0.01878	0.00151	0.06720
$t = 6$	0.01994	0.00161	0.07132
$t = 7$	0.02079	0.00168	0.07438
$t = 8$	0.02144	0.00173	0.07668
$t = 9$	0.02193	0.00177	0.07844
$t = 10$	0.02231	0.00179	0.07981

H FIGURES

In this section, we provide some empirical evidence to support the results in Section 4: **Numerical Experiments**. We will show how the trajectories of the non-convex deep linear neural networks are

related to a convex optimization problem for GD under different initialization schemes. Consider the following procedures for plots of the logarithm of loss as a function of number of iterations:

a) We choose $X \in \mathbb{R}^{128\times 1000}$ and $W_* \in \mathbb{R}^{10\times 128}$ and set $Y = W_* X + \varepsilon$, where the entries in X, W_* and ε are drawn i.i.d. from $N(0, 1)$.

b) We consider the loss function $\frac{1}{2}\left\|a_N W_{N:1} X - Y\right\|_F^2$.

c) For the given linear networks, we apply the Gaussian initialization and the one peak random orthogonal projections and embeddings initialization, which are denoted as $W_j(0), 1 \leq j \leq N$.

d) For the convex optimization problem (1), we set the initialization to be $W(0) = a_N W_N(0) \cdots W_1(0)$.

e) We set the learning rate $\eta = \frac{n_N}{N\cdot\|X\|^2}$ and $\eta_* = \frac{N}{n_N}\eta$ for the deep linear neural networks and the convex problem, respectively.

f) We draw the loss function through 25 iterations.

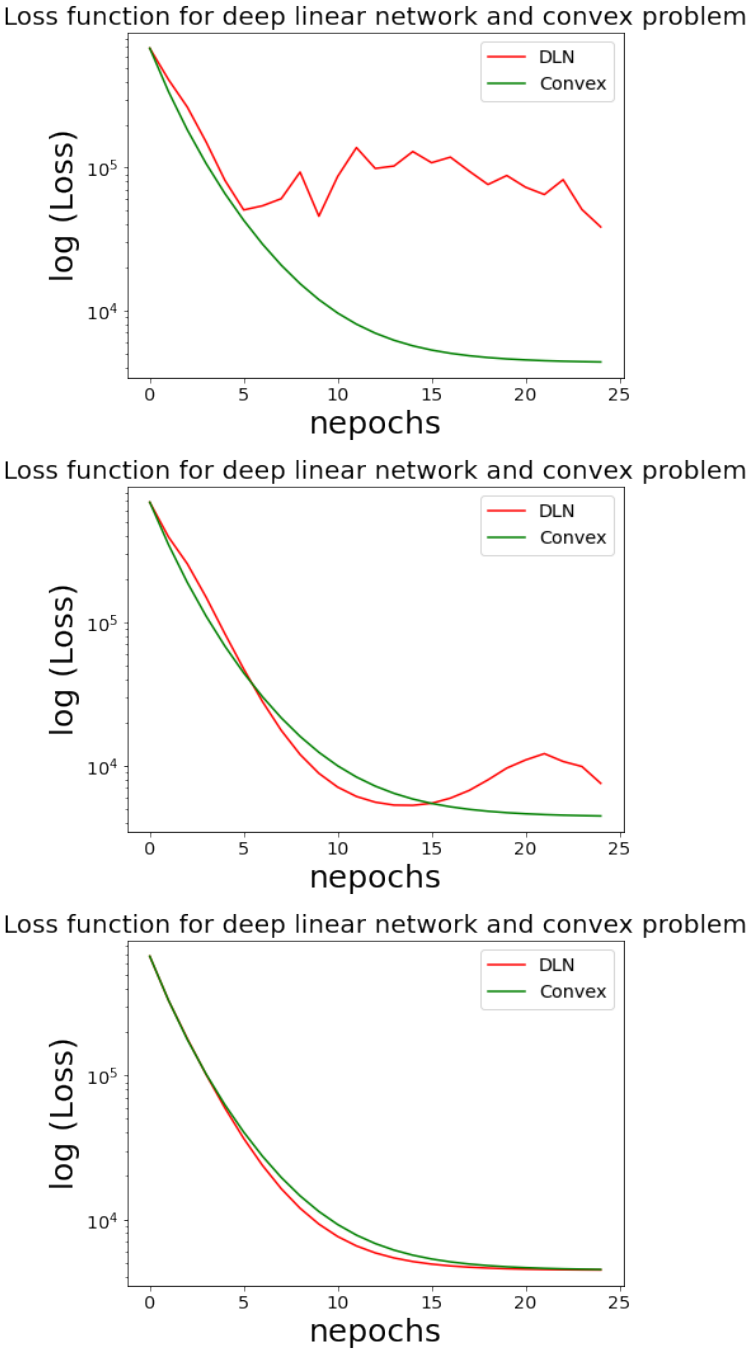

Figure 1: Plot of Loss as a function of number of iterations with $n_1 = n_2 = n_3 = 128$ (First), 200 (Second), 2000 (Third) for Gaussian initialization, respectively.

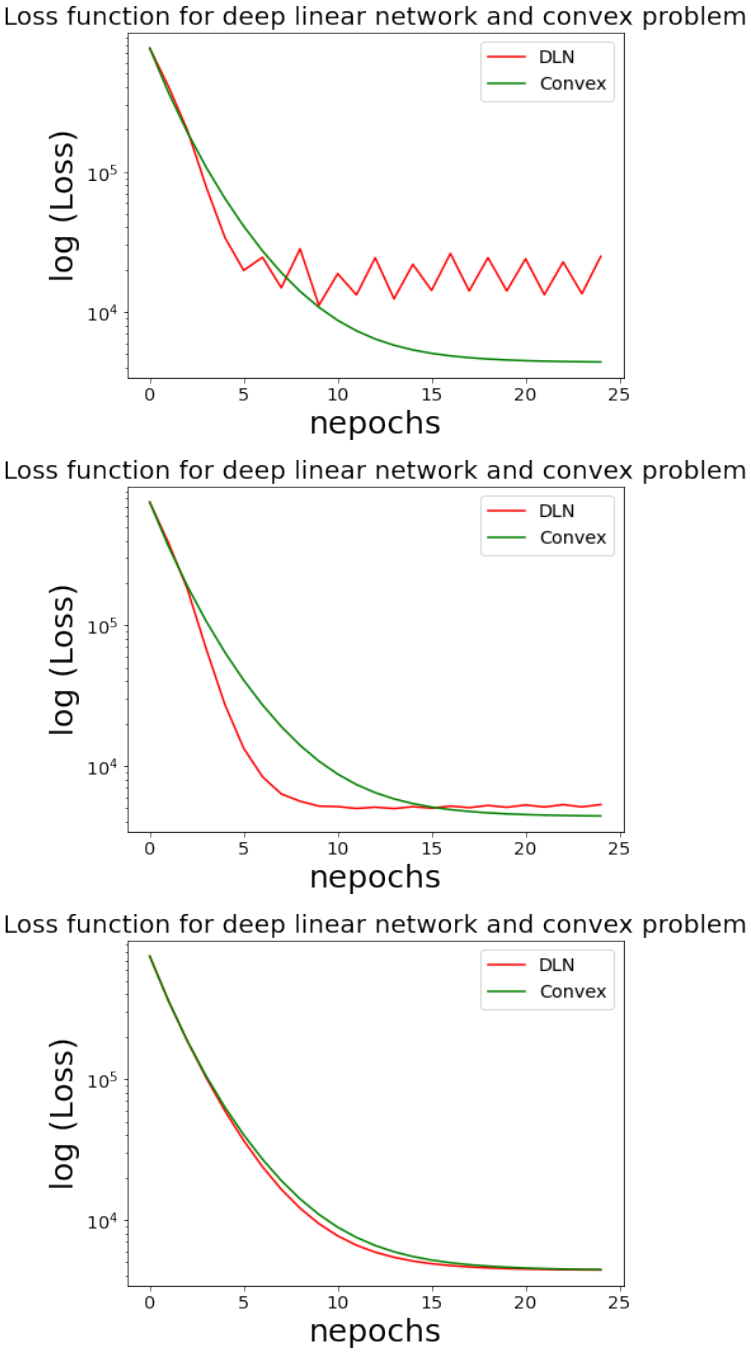

Figure 2: Plot of Loss as a function of number of iterations with $n_1 = n_2 = n_3 = 128$ (First), 200 (Second), 5000 (Third) for Orthogonal initialization, respectively.