Understanding Generalization and Optimization Performance for Deep CNNs (Supplementary File)

Pan Zhou* Jiashi Feng*
* National University of Singapore, Singapore
pzhou@u.nus.edu elefjia@nus.edu.sg

A Structure of This Document

This document gives some other necessary notations and preliminaries for our analysis in Sec. B.1 and provides auxiliary lemmas in Sec. B.2. Then in Sec. C, we present the technical lemmas for proving our final results and their proofs. Next, in Sec. D, we utilize these technical lemmas to prove our desired results. Finally, we give the proofs of other auxiliary lemmas in Sec. E.

As for the results in manuscript, the proofs of Lemma 1 and Theorem 1 in Sec. 3.1 in the manuscript are respectively provided in Sec. D.1 and Sec. D.2. As for the results in Sec. 4 in the manuscript, Sec. D.3 and D.4 present the proofs of Theorem 2 and Corollary 1, respectively. Finally, we respectively introduce the proofs of Theorem 3 and Corollary 2 in Sec. D.5 and D.6.

B Notations and Preliminary Tools

Beyond the notations introduced in the manuscript, we need some other notations used in this document. Then we introduce several lemmas that will be used later.

B.1 Notations

Throughout this document, we use $\langle\cdot,\cdot\rangle$ to denote the inner product and use $\widetilde{\circledast}$ to denote the convolution operation with stride 1. $\boldsymbol{A}\otimes\boldsymbol{C}$ denotes the Kronecker product between $\boldsymbol{A}$ and $\boldsymbol{C}$. Note that $\boldsymbol{A}$ and $\boldsymbol{C}$ in $\boldsymbol{A}\otimes\boldsymbol{C}$ can be matrices or vectors. For a matrix $\boldsymbol{A}\in\mathbb{R}^{n_1\times n_2}$, we use $\|\boldsymbol{A}\|_F=\sqrt{\sum_{i,j}\boldsymbol{A}_{ij}^2}$ to denote its Frobenius norm, where $\boldsymbol{A}_{ij}$ is the (i,j)-th entry of $\boldsymbol{A}$. We use $\|\boldsymbol{A}\|_{\text{op}}=\max_i|\lambda_i(\boldsymbol{A})|$ to denote the operation norm of a matrix $\boldsymbol{A}\in\mathbb{R}^{n_1\times n_1}$, where $\lambda_i(\boldsymbol{A})$ denotes the i-th eigenvalue of the matrix $\boldsymbol{A}$. For a 3-way tensor $\boldsymbol{A}\in\mathbb{R}^{s\times t\times q}$, its operation norm is computed as

$$\|\boldsymbol{A}\|_{\text{op}}=\sup_{\|\boldsymbol{\lambda}\|_2\leq 1}\left\langle\boldsymbol{\lambda}^{\otimes^3},\boldsymbol{A}\right\rangle=\sum_{i,j,k}\boldsymbol{A}_{ijk}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_j\boldsymbol{\lambda}_k,$$

where $\boldsymbol{A}_{ijk}$ denotes the (i,j,k)-th entry of $\boldsymbol{A}$.

For brevity, in this document we use $f(\boldsymbol{w},\boldsymbol{D})$ to denote $f(g(\boldsymbol{w};\boldsymbol{D}),\boldsymbol{y})$ in the manuscript. Let $\boldsymbol{w}_{(i)}=(\boldsymbol{w}_{(i)}^1;\cdots;\boldsymbol{w}_{(i)}^{d_i})\in\mathbb{R}^{{k_i}^2d_{i-1}d_i}$ $(i=1,\cdots,l)$ be the parameter of the i-th layer where $\boldsymbol{w}_{(i)}^k=\mathsf{vec}\left(\boldsymbol{W}_{(i)}^k\right)\in\mathbb{R}^{{k_i}^2d_{i-1}}$ is the vectorization of $\boldsymbol{W}_{(i)}^k$. Similarly, let $\boldsymbol{w}_{(l+1)}=\mathsf{vec}\left(\boldsymbol{W}_{(l+1)}\right)\in\mathbb{R}^{r_lc_ld_ld_{l+1}}$. Then, we further define $\boldsymbol{w}=(\boldsymbol{w}_{(1)},\cdots,\boldsymbol{w}_{(l)},\boldsymbol{w}_{(l+1)})\in\mathbb{R}^{r_lc_ld_ld_{l+1}+\sum_{i=1}^l{k_i}^2d_{i-1}d_i}$ which contains all the parameter in the network. Here we use $\boldsymbol{W}_{(i)}^k$ to denote the k-th kernel in the i-th convolutional layer. For brevity, let $\boldsymbol{W}_{(i)}^{k,j}$ denotes the j-th slice of $\boldsymbol{W}_{(i)}^k$, *i.e.* $\boldsymbol{W}_{(i)}^{k,j}=\boldsymbol{W}_{(i)}^k(:,:,j)$.

For a matrix $\boldsymbol{M}\in\mathbb{R}^{s\times t}$, $\widehat{\boldsymbol{M}}$ denotes the matrix which is obtained by rotating the matrix $\boldsymbol{M}$ by 180 degrees. Then we use $\boldsymbol{\delta}_i$ to denote the gradient of $f(\boldsymbol{w},\boldsymbol{D})$ *w.r.t.* $\boldsymbol{X}_{(i)}$:

$$\boldsymbol{\delta}_i=\nabla_{\boldsymbol{X}_{(i)}}f(\boldsymbol{w},\boldsymbol{D})\in\mathbb{R}^{r_i\times c_i\times d_i},\quad(i=1,\cdots,l),$$

Based on $\boldsymbol{\delta}_i$, we further define $\widetilde{\boldsymbol{\delta}}_i \in \mathbb{R}^{(\tilde{r}_{i-1}-k_i+1)\times(\tilde{c}_{i-1}-k_i+1)\times d_i}$. Each slice $\widetilde{\boldsymbol{\delta}}_{i+1}^k$ can be computed as follows. Firstly, let $\widetilde{\boldsymbol{\delta}}_{i+1}^k = \boldsymbol{\delta}_{i+1}^k$. Then, we pad zeros of $s_i - 1$ rows between the neighboring rows in $\widetilde{\boldsymbol{\delta}}_{i+1}^k$ and similarly we pad zeros of $s_i - 1$ columns between the neighboring columns in $\widetilde{\boldsymbol{\delta}}_{i+1}^k$. Accordingly, the size of $\widetilde{\boldsymbol{\delta}}_{i+1}^k$ is $(s_i(r_i-1)+1)\times(s_i(c_i-1)+1)$. Finally, we pad zeros of width $k_i - 1$ around $\widetilde{\boldsymbol{\delta}}_{i+1}^k$ to obtain new $\widetilde{\boldsymbol{\delta}}_{i+1}^k \in \mathbb{R}^{(s_i(r_i-1)+2k_i-1)\times(s_i(c_i-1)+2k_i-1)}$. Note that since $r_{i+1} = (\tilde{r}_i - k_{i+1})/s_{i+1} + 1$ and $r_{i+1} = (\tilde{r}_i - k_{i+1})/s_{i+1} + 1$, we have $s_i(r_i-1) + 2k_i - 1 = \tilde{r}_{i-1} - k_i + 1$ and $s_i(c_i-1) + 2k_i - 1 = \tilde{c}_{i-1} - k_i + 1$.

Then we define four operators $\mathsf{G}(\cdot)$, $\mathsf{Q}(\cdot)$, $\mathsf{up}(\cdot)$ and $\mathsf{vec}(\cdot)$, which are useful for explaining the following analysis.

Operation $\mathsf{G}(\cdot)$: It maps an arbitrary vector $\boldsymbol{z} \in \mathbb{R}^d$ into a diagonal matrix $\mathsf{G}(\boldsymbol{z}) \in \mathbb{R}^{d\times d}$ with its i-th diagonal entry equal to $\sigma_1(\boldsymbol{z}_i)(1-\sigma_1(\boldsymbol{z}_i))$ in which $\boldsymbol{z}_i$ denotes the i-th entry of $\boldsymbol{z}$.

Operation $\mathsf{Q}(\cdot)$: it maps a vector $\boldsymbol{z} \in \mathbb{R}^d$ into a matrix of size $d^2 \times d$ whose $((i-1)d+i, i)$ $(i = 1, \cdots, d)$ entry equal to $\sigma_1(\boldsymbol{z}_i)(1-\sigma_1(\boldsymbol{z}_i))(1-2\sigma_1(\boldsymbol{z}_i))$ and rest entries are all 0.

Operation $\mathsf{up}(\cdot)$: $\mathsf{up}(\boldsymbol{M})$ represents conducting upsampling on $\boldsymbol{M} \in \mathbb{R}^{s\times t\times q}$. Let $\boldsymbol{N} = \mathsf{up}(\boldsymbol{M}) \in \mathbb{R}^{ps\times pt\times q}$. Specifically, for each slice $\boldsymbol{N}(:,:,i)$ $(i = 1, \cdots, q)$, we have $\boldsymbol{N}(:,:,i) = \mathsf{up}(\boldsymbol{M}(:,:,i))$. It actually upsamples each entry $\boldsymbol{M}(g,h,i)$ into a matrix of p^2 same entries $\frac{1}{p^2}\boldsymbol{M}(g,h,i)$.

Operation $\mathsf{vec}(\cdot)$: For a matrix $\boldsymbol{A} \in \mathbb{R}^{s\times t}$, $\mathsf{vec}(\boldsymbol{A})$ is defined as $\mathsf{vec}(\boldsymbol{A}) = (\boldsymbol{A}(:,1); \cdots; \boldsymbol{A}(:,t)) \in \mathbb{R}^{st}$ that vectorizes $\boldsymbol{A} \in \mathbb{R}^{s\times t}$ along its columns. If $\boldsymbol{A} \in \mathbb{R}^{t\times s\times q}$ is a 3-way tensor, $\mathsf{vec}(\boldsymbol{A}) = [\mathsf{vec}(\boldsymbol{A}(:,:,1)); \mathsf{vec}(\boldsymbol{A}(:,:,2)), \cdots, \mathsf{vec}(\boldsymbol{A}(:,:,q))] \in \mathbb{R}^{stq}$.

B.2 Auxiliary Lemmas

We first introduce Lemmas 2 and 3 which are respectively used for bounding the ℓ_2-norm of a vector and the operation norm of a matrix. Then we introduce Lemmas 4 and 5 which discuss the topology of functions. In Lemma 6, we introduce the Hoeffding's inequality which provides an upper bound on the probability that the sum of independent random variables deviates from its expected value. In Lemma 7, we provide the covering number of an ϵ-net for a low-rank matrix. Finally, several commonly used inequalities are presented in Lemma 8 for our analysis.

Lemma 2. *(Vershynin, 2012) For any vector $\boldsymbol{x} \in \mathbb{R}^d$, its ℓ_2-norm can be bounded as*

$$\|\boldsymbol{x}\|_2 \leq \frac{1}{1-\epsilon} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}_\epsilon} \langle \boldsymbol{\lambda}, \boldsymbol{x}\rangle .$$

where $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k_w}\}$ be an ϵ-covering net of $\mathsf{B}^d(1)$.

Lemma 3. *(Vershynin, 2012) For any symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, its operator norm can be bounded as*

$$\|\boldsymbol{X}\|_{op} \leq \frac{1}{1-2\epsilon} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}_\epsilon} |\langle \boldsymbol{\lambda}, \boldsymbol{X}\boldsymbol{\lambda}\rangle| .$$

where $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k_w}\}$ be an ϵ-covering net of $\mathsf{B}^d(1)$.

Lemma 4. *(Mei et al., 2017) Let $D \subseteq \mathbb{R}^d$ be a compact set with a C^2 boundary ∂D, and $f, g : A \to \mathbb{R}$ be C^2 functions defined on an open set A, with $D \subseteq A$. Assume that for all $\boldsymbol{w} \in \partial D$ and all $t \in [0,1]$, $t\nabla f(\boldsymbol{w}) + (1-t)\nabla g(\boldsymbol{w}) \neq \boldsymbol{0}$. Finally, assume that the Hessian $\nabla^2 f(\boldsymbol{w})$ is non-degenerate and has index equal to r for all $\boldsymbol{w} \in D$. Then the following properties hold:*

(1) If g has no critical point in D, then f has no critical point in D.

(2) If g has a unique critical point $\boldsymbol{w}$ in D that is non-degenerate with an index of r, then f also has a unique critical point $\boldsymbol{w}'$ in D with the index equal to r.

Lemma 5. *(Mei et al., 2017) Suppose that $F(\boldsymbol{w}) : \Theta \to \mathbb{R}$ is a C^2 function where $\boldsymbol{w} \in \Theta$. Assume that $\{\boldsymbol{w}_{(1)}, \ldots, \boldsymbol{w}_{(m)}\}$ is its non-degenerate critical points and let $D = \{\boldsymbol{w} \in \Theta : \|\nabla F(\boldsymbol{w})\|_2 < \epsilon$ and $\inf_i \left|\lambda_i\left(\nabla^2 F(\boldsymbol{w})\right)\right| \geq \zeta\}$. Then D can be decomposed into (at most) countably components, with each component containing either exactly one critical point, or no critical point. Concretely, there exist disjoint open sets $\{D_k\}_{k\in\mathbb{N}}$, with D_k possibly empty for $k \geq m+1$, such that*

$$D = \cup_{k=1}^{\infty} D_k .$$

Furthermore, $\boldsymbol{w}_{(k)} \in D_k$ for $1 \leq k \leq m$ and each D_i, $k \geq m+1$ contains no stationary points.

Lemma 6. *(Hoeffding, 1963) Let $x_1, \cdots, x_n$ be independent random variables where x_i is bounded by the interval $[a_i, b_i]$. Suppose $s_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then the following properties hold:*

$$\mathbb{P}\left(s_n - \mathbb{E}s_n \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

where t is an arbitrary positive value.

Lemma 7. *(Candes & Plan, 2009) Let $\mathbb{S}_r = \{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\boldsymbol{X}) \leq r, \|\boldsymbol{X}\|_F = b\}$. Then there exists an ϵ-net $\widetilde{\mathbb{S}}_r$ for the Frobenius norm obeying*

$$|\widetilde{\mathbb{S}}_r| \leq \left(\frac{9b}{\epsilon}\right)^{r(n_1+n_2+1)}.$$

Then we give a lemma to summarize the properties of $\mathsf{G}(\cdot)$, $\mathsf{Q}(\cdot)$ and $\mathsf{up}(\cdot)$ defined in Section B.1, the convolutional operation $\circledast$ defined in manuscript.

Lemma 8. *For $\mathsf{G}(\cdot)$, $\mathsf{Q}(\cdot)$ and $\mathsf{up}(\cdot)$ defined in Section B.1, the convolutional operation $\circledast$ defined in manuscript, we have the following properties:*

(1) For arbitrary vector $\boldsymbol{z}$, and arbitrary matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ of proper sizes, we have

$$\|\mathsf{G}(\boldsymbol{z})\boldsymbol{M}\|_F^2 \leq \frac{1}{16}\|\boldsymbol{M}\|_F^2 \quad \textit{and} \quad \|\boldsymbol{N}\mathsf{G}(\boldsymbol{z})\|_F^2 \leq \frac{1}{16}\|\boldsymbol{N}\|_F^2.$$

(2) For arbitrary vector $\boldsymbol{z}$, and arbitrary matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ of proper sizes, we have

$$\|\mathsf{Q}(\boldsymbol{z})\boldsymbol{M}\|_F^2 \leq \frac{2^6}{3^8}\|\boldsymbol{M}\|_F^2 \quad \textit{and} \quad \|\boldsymbol{N}\mathsf{Q}(\boldsymbol{z})\|_F^2 \leq \frac{2^6}{3^8}\|\boldsymbol{N}\|_F^2.$$

(3) For any tensor $\boldsymbol{M} \in \mathbb{R}^{s\times t\times q}$, we have

$$\|\mathsf{up}(\boldsymbol{M})\|_F^2 \leq \frac{1}{p^2}\|\boldsymbol{M}\|_F^2.$$

(4) For any kernel $\boldsymbol{W} \in \mathbb{R}^{k_i\times k_i\times d_i}$ and $\widetilde{\boldsymbol{\delta}}_{i+1} \in \mathbb{R}^{(\tilde{r}_{i-1}-k_i+1)\times(\tilde{c}_{i-1}-k_i+1)\times d_i}$ defined in Sec. B.1, then we have

$$\|\widetilde{\boldsymbol{\delta}}_{i+1}\widetilde{\circledast}\boldsymbol{W}\|_F^2 \leq (k_i - s_i + 1)^2\|\boldsymbol{W}\|_F^2\|\widetilde{\boldsymbol{\delta}}_{i+1}\|_F^2.$$

(5) For softmax activation function σ_2, we can bound the norm of difference between output $\boldsymbol{v}$ and its corresponding ont-hot label as

$$0 \leq \|\boldsymbol{v} - \boldsymbol{y}\|_2^2 \leq 2.$$

It should be pointed out that we defer the proof of Lemma 8 to Sec. E.

C Technical Lemmas and Their Proofs

Here we present the key lemmas and theorems for proving our desired results. For brevity, in this document we use $f(\boldsymbol{w}, \boldsymbol{D})$ to denote $f(g(\boldsymbol{w}; \boldsymbol{D}), \boldsymbol{y})$ in the manuscript.

Lemma 9. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Then the gradient of $f(\boldsymbol{w}, \boldsymbol{D})$ with respect to $\boldsymbol{w}_{(j)}$ can be written as*

$$\begin{aligned}
\nabla_{\boldsymbol{W}_{(l+1)}} f(\boldsymbol{w}, \boldsymbol{D}) =& \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y})\boldsymbol{z}_{(l)}^T \in \mathbb{R}^{d_{l+1}\times\tilde{r}_l\tilde{c}_l d_l}, \\
\nabla_{\boldsymbol{W}_{(i)}^{k,j}} f(\boldsymbol{w}, \boldsymbol{D}) =& \boldsymbol{Z}_{(i-1)}^j \widetilde{\circledast} \widetilde{\boldsymbol{\delta}}_i^k \in \mathbb{R}^{k_i\times k_i}, \ (j = 1, \cdots, d_{i-1};\ k = 1, \cdots, d_i;\ i = 1, \cdots, l),
\end{aligned}$$

where $\boldsymbol{\delta}_i^k$ is the k-slice (i.e. $\boldsymbol{\delta}_i(:,:,k)$) of $\boldsymbol{\delta}_i$ which is defined as

$$\boldsymbol{\delta}_i = \nabla_{\boldsymbol{X}_{(i)}} f(\boldsymbol{w}, \boldsymbol{D}) \in \mathbb{R}^{r_i \times c_i \times d_i}, \quad (i = 1, \cdots, l),$$

and further satisfies

$$\boldsymbol{\delta}_i^j = \mathsf{up}\left(\sum_{k=1}^{d_{i+1}} \widetilde{\boldsymbol{\delta}}_{i+1}^k \widetilde{\circledast} \widehat{\boldsymbol{W}}_{(i+1)}^{k,j}\right) \odot \sigma_1'(\boldsymbol{X}_{(i)}^j) \in \mathbb{R}^{r_i \times c_i}, \quad (j = 1, \cdots, d_i;\ i = 1, \cdots, l-1),$$

where the matrix $\widehat{\boldsymbol{W}}_{(i+1)}^{k,j} \in \mathbb{R}^{k_{i+1} \times k_{i+1}}$ is obtained by rotating $\boldsymbol{W}_{(i+1)}^{k,j}$ with 180 degrees. Also, $\boldsymbol{\delta}_l$ is computed as follows:

$$\nabla_{\boldsymbol{u}} f(\boldsymbol{w}, \boldsymbol{D}) = \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y}) \in \mathbb{R}^{d_{l+1}}, \qquad \nabla_{\boldsymbol{z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D}) = (\boldsymbol{W}_{(l+1)})^T \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y}) \in \mathbb{R}^{\tilde{r}_l \tilde{c}_l d_l},$$

$$\nabla_{\boldsymbol{Z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D}) = \mathsf{reshape}\left(\nabla_{\boldsymbol{z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D})\right) \in \mathbb{R}^{\tilde{r}_l \times \tilde{c}_l \times d_l},\ \boldsymbol{\delta}_l = \sigma_1'(\boldsymbol{X}_{(l)}) \odot \mathsf{up}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{Z}_{(l)}}\right) \in \mathbb{R}^{r_i \times c_i \times d_i}.$$

where $\boldsymbol{S} = \mathsf{G}(\boldsymbol{u})$.

Proof. We use chain rule to compute the gradient of $f(\boldsymbol{w}, \boldsymbol{D})$ with respect to $\boldsymbol{Z}_{(i)}$. We first compute several basis gradient. According to the relationship between $\boldsymbol{X}_{(i)}, \boldsymbol{Y}_{(i)}, \boldsymbol{Z}_{(i)}$ and $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{D})$, we have

$$\begin{aligned}
&\nabla_{\boldsymbol{u}} f(\boldsymbol{w}, \boldsymbol{D}) = \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y}) \in \mathbb{R}^{d_{l+1}}, \\
&\nabla_{\boldsymbol{z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D}) = (\boldsymbol{W}_{(l+1)})^T \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y}) \in \mathbb{R}^{\tilde{r}_l \tilde{c}_l d_l}, \\
&\nabla_{\boldsymbol{Z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D}) = \mathsf{reshape}\left(\nabla_{\boldsymbol{z}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D})\right) \in \mathbb{R}^{\tilde{r}_l \times \tilde{c}_l \times d_l}, \\
&\nabla_{\boldsymbol{X}_{(l)}} f(\boldsymbol{w}, \boldsymbol{D}) = \frac{\partial \boldsymbol{Y}_{(l)}}{\partial \boldsymbol{X}_{(l)}} \frac{\partial \boldsymbol{Z}_{(l)}}{\partial \boldsymbol{Y}_{(l)}} \frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{Z}_{(l)}} = \sigma_1'(\boldsymbol{X}_{(l)}) \odot \mathsf{up}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{Z}_{(l)}}\right) \triangleq \boldsymbol{\delta}_l \in \mathbb{R}^{r_i \times c_i \times d_i}.
\end{aligned}$$

Then we can further obtain

$$\boldsymbol{\delta}_i^j = \mathsf{up}\left(\sum_{k=1}^{d_{i+1}} \widetilde{\boldsymbol{\delta}}_{i+1}^k \widetilde{\circledast} \widehat{\boldsymbol{W}}_{(i+1)}^{k,j}\right) \odot \sigma_1'(\boldsymbol{X}_{(i)}^j) \in \mathbb{R}^{r_i \times c_i}, \quad (j = 1, \cdots, d_i;\ i = 1, \cdots, l-1).$$

where $\widehat{\boldsymbol{W}}_{(i)}^{k,j}$ denotes the j-th slice $\widehat{\boldsymbol{W}}_{(i)}^k(:,:j)$ of $\widehat{\boldsymbol{W}}_{(i)}^k$. Note, we clockwise rotate the matrix $\boldsymbol{W}_{(i+1)}^{k,j}$ by 180 degrees to obtain $\widehat{\boldsymbol{W}}_{(i+1)}^{k,j}$. Finally, we can compute the gradient *w.r.t.* $\boldsymbol{W}_{(l+1)}$ and $\boldsymbol{W}^{(i)}$ $(i = 1, \cdots, l)$:

$$\begin{aligned}
\nabla_{\boldsymbol{W}_{(l+1)}} f(\boldsymbol{w}, \boldsymbol{D}) &= \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y}) \boldsymbol{z}_{(l)}^T \in \mathbb{R}^{d_{l+1} \times \tilde{r}_l \tilde{c}_l d_l}, \\
\nabla_{\boldsymbol{W}_{(i)}^{k,j}} f(\boldsymbol{w}, \boldsymbol{D}) &= \boldsymbol{X}_{(i-1)}^j \widetilde{\circledast} \widetilde{\boldsymbol{\delta}}_i^k \in \mathbb{R}^{k_i \times k_i},\ (j = 1, \cdots, d_{i-1};\ k = 1, \cdots, d_i;\ i = 1, \cdots, l).
\end{aligned}$$

The proof is completed. □

Lemma 10. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Then the gradient of $f(\boldsymbol{w}, \boldsymbol{D})$ with respect to $\boldsymbol{w}_{(j)}$ can be written as*

$$\|\boldsymbol{\delta}_l\|_F^2 \leq \frac{\vartheta}{16p^2} {b_{l+1}}^2, \quad \|\boldsymbol{\delta}_i\|_F^2 \leq \frac{d_{i+1} {b_{i+1}}^2 (k_{i+1} - s_{i+1} + 1)^2}{16p^2} \left\|\boldsymbol{\delta}_{i+1}\right\|_F^2, \quad \|\boldsymbol{\delta}_i\|_F^2 \leq \frac{\vartheta {b_{l+1}}^2}{16p^2} \prod_{s=i+1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2},$$

where $\vartheta = 1/8$.

Proof. We first bound $\boldsymbol{\delta}_l$. By Lemma 9, we have

$$\begin{aligned}
\|\boldsymbol{\delta}_l\|_F^2 &= \left\|\sigma_1'(\boldsymbol{X}_{(l)}) \odot \mathsf{up}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{Z}_{(l)}}\right)\right\|_F^2 \overset{①}{\leq} \frac{1}{16p^2} \left\|\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{Z}_{(l)}}\right\|_F^2 = \frac{1}{16p^2} \left\|\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{z}_{(l)}}\right\|_2^2 \\
&\leq \frac{1}{16p^2} \left\|(\boldsymbol{W}_{(l+1)})^T \boldsymbol{S}(\boldsymbol{v} - \boldsymbol{y})\right\|_2^2 \overset{②}{\leq} \frac{\vartheta}{16p^2} {b_{l+1}}^2,
\end{aligned}$$

where ① holds since the values of the entries in the tensor $\sigma_1'(\boldsymbol{X}_{(l)}) \in \mathbb{R}^{r_l \times c_l \times d_l}$ belong to $[0, 1/4]$, and the $\mathsf{up}(\cdot)$ operation does repeat one entry into p^2 entries but the entry value becomes $1/p^2$ of the original entry value. ② holds since we have $\|\boldsymbol{S}(\boldsymbol{v}-\boldsymbol{y})\|_2^2 \leq 1/8$ in Lemma 8.

Also by Lemma 9, we can further bound $\|\boldsymbol{\delta}_i^j\|_F^2$ as follows:

$$\begin{aligned}
\|\boldsymbol{\delta}_i\|_F^2 =& \sum_{j=1}^{d_i} \left\|\boldsymbol{\delta}_i^j\right\|_F^2 = \sum_{j=1}^{d_i} \left\| \mathsf{up}\left(\sum_{k=1}^{d_{i+1}} \widetilde{\boldsymbol{\delta}}_{i+1}^k \widetilde{\circledast} \widehat{\boldsymbol{W}}_{(i+1)}^{k,j}\right) \odot \sigma_1'(\boldsymbol{X}_{(i)}^j) \right\|_F^2 \leq \frac{1}{16p^2} \sum_{j=1}^{d_i} \left\| \sum_{k=1}^{d_{i+1}} \widetilde{\boldsymbol{\delta}}_{i+1}^k \widetilde{\circledast} \widehat{\boldsymbol{W}}_{(i+1)}^{k,j} \right\|_F^2 \\
\leq& \frac{d_{i+1}}{16p^2} \sum_{j=1}^{d_i} \sum_{k=1}^{d_{i+1}} \left\| \widetilde{\boldsymbol{\delta}}_{i+1}^k \widetilde{\circledast} \widehat{\boldsymbol{W}}_{(i+1)}^{k,j} \right\|_F^2 \leq \frac{d_{i+1}(k_{i+1}-s_{i+1}+1)^2}{16p^2} \sum_{j=1}^{d_i} \sum_{k=1}^{d_{i+1}} \left\| \widetilde{\boldsymbol{\delta}}_{i+1}^k \right\|_F^2 \left\| \widehat{\boldsymbol{W}}_{(i+1)}^{k,j} \right\|_F^2 \\
\overset{①}{=}& \frac{d_{i+1}(k_{i+1}-s_{i+1}+1)^2}{16p^2} \sum_{k=1}^{d_{i+1}} \left\|\boldsymbol{\delta}_{i+1}^k\right\|_F^2 \left\|\boldsymbol{W}_{(i+1)}^k\right\|_F^2 \leq \frac{d_{i+1} {b_{i+1}}^2 (k_{i+1}-s_{i+1}+1)^2}{16p^2} \left\|\boldsymbol{\delta}_{i+1}\right\|_F^2,
\end{aligned}$$

where $\widehat{\boldsymbol{W}}_{(i+1)}^{k,j}$ denotes the j-th slice $\widehat{\boldsymbol{W}}_{(i+1)}^k(:,:,j)$ of the tensor $\widehat{\boldsymbol{W}}_{(i+1)}^k$. ① holds since we rotate the matrix $\boldsymbol{W}_{(i+1)}^{k,j}$ by 180 degrees to obtain $\widehat{\boldsymbol{W}}_{(i+1)}^k$, indicating $\|\boldsymbol{W}_{(i+1)}^{k,j}\|_F^2 = \|\widehat{\boldsymbol{W}}_{(i+1)}^{k,j}\|_F^2$ and $\sum_{j=1}^{d_i} \left\|\boldsymbol{W}_{(i+1)}^{k,j}\right\|_F^2 = \left\|\boldsymbol{W}_{(i+1)}^k\right\|_F^2 \leq {b_{i+1}}^2$. Accordingly, the above inequality gives

$$\begin{aligned}
\|\boldsymbol{\delta}_i\|_F^2 \leq& \frac{d_{i+1} {b_{i+1}}^2 (k_{i+1}-s_{i+1}+1)^2}{16p^2} \left\|\boldsymbol{\delta}_{i+1}\right\|_F^2 \leq \cdots \leq \|\boldsymbol{\delta}_l\|_F^2 \prod_{s=i+1}^{l} \frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2} \\
\leq& \frac{\vartheta {b_{l+1}}^2}{16p^2} \prod_{s=i+1}^{l} \frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2}.
\end{aligned}$$

The proof is completed. □

Lemma 11. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Then the gradient of $f(\boldsymbol{w}, \boldsymbol{D})$ with respect to $\boldsymbol{W}^{(l+1)}$ and $\boldsymbol{w}$ can be respectively bounded as follows:*

$$\left\|\nabla_{\boldsymbol{W}_{(l+1)}} f(\boldsymbol{w}, \boldsymbol{D})\right\|_F^2 \leq \vartheta \tilde{r}_l \tilde{c}_l d_l \quad \textit{and} \quad \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D})\|_2^2 \leq \beta^2,$$

where $\vartheta = 1/8$ and $\beta \triangleq \left[\vartheta \tilde{r}_l \tilde{c}_l d_l + \sum_{i=1}^{l} \frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1} c_{i-1} \prod_{s=i}^{l} \frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2}\right]^{1/2}$.

Proof. By utilizing Lemma 10, we can bound

$$\begin{aligned}
\sum_{i=1}^{l} \left\|\nabla_{\boldsymbol{w}_{(i)}} f(\boldsymbol{w}, \boldsymbol{D})\right\|_F^2 =& \sum_{i=1}^{l} \sum_{k=1}^{d_i} \sum_{j=1}^{d_{i-1}} \left\|\nabla_{\boldsymbol{W}_{(i)}^{k,j}} f(\boldsymbol{w}, \boldsymbol{D})\right\|_F^2 = \sum_{i=1}^{l} \sum_{k=1}^{d_i} \sum_{j=1}^{d_{i-1}} \left\|\boldsymbol{Z}_{(i-1)}^j \widetilde{\circledast} \boldsymbol{\delta}_i^k\right\|_F^2 \\
\overset{①}{\leq}& \sum_{i=1}^{l} \sum_{k=1}^{d_i} \sum_{j=1}^{d_{i-1}} (k_i - s_i + 1)^2 \left\|\boldsymbol{Z}_{(i-1)}^j\right\|_F^2 \left\|\boldsymbol{\delta}_i^k\right\|_F^2 \\
\overset{②}{\leq}& \sum_{i=1}^{l} \sum_{k=1}^{d_i} \sum_{j=1}^{d_{i-1}} \tilde{r}_{i-1} \tilde{c}_{i-1} (k_i - s_i + 1)^2 \left\|\boldsymbol{\delta}_i^k\right\|_F^2 \\
\leq& \sum_{i=1}^{l} \tilde{r}_{i-1} \tilde{c}_{i-1} d_{i-1} (k_i - s_i + 1)^2 \left\|\boldsymbol{\delta}_i\right\|_F^2 \\
\leq& \sum_{i=1}^{l} \tilde{r}_{i-1} \tilde{c}_{i-1} d_{i-1} (k_i - s_i + 1)^2 \frac{\vartheta {b_{l+1}}^2}{16p^2} \prod_{s=i+1}^{l} \frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2} \\
=& \sum_{i=1}^{l} \frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1} c_{i-1} \prod_{s=i}^{l} \frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2},
\end{aligned}$$

where ① holds since $\left\|\boldsymbol{Z}_{(i-1)}^{j}\widetilde{\circledast}\boldsymbol{\delta}_i^k\right\|_F^2 \leq (k_i-s_i+1)^2\left\|\boldsymbol{Z}_{(i-1)}^{j}\right\|_F^2\left\|\boldsymbol{\delta}_i^k\right\|_F^2$; and ② holds since the values of entries in $\boldsymbol{Z}_{(i-1)}^{j}$ belong to $[0,1]$.

On the other hand, we can bound

$$\left\|\nabla_{\boldsymbol{W}_{(l+1)}} f(\boldsymbol{w},\boldsymbol{D})\right\|_F^2 = \left\|\boldsymbol{S}(\boldsymbol{v}-\boldsymbol{y})\boldsymbol{z}_{(l)}^T\right\|_F^2 \leq \vartheta\tilde{r}_l\tilde{c}_l d_l.$$

So we can bound the ℓ_2 norm of the gradient as follows:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w},\boldsymbol{D})\right\|_F^2 =& \left\|\nabla_{\boldsymbol{W}_{(l+1)}} f(\boldsymbol{w},\boldsymbol{x})\right\|_F^2 + \sum_{i=1}^{l}\left\|\nabla_{\boldsymbol{w}_{(i)}} f(\boldsymbol{w},\boldsymbol{D})\right\|_F^2 \\ \leq& \vartheta\tilde{r}_l\tilde{c}_l d_l + \sum_{i=1}^{l}\frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1}c_{i-1}\prod_{s=i}^{l}\frac{d_s {b_s}^2(k_s-s_s+1)^2}{16p^2}.\end{aligned}$$

The proof is completed. □

Lemma 12. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w},\boldsymbol{D})$ is squared loss. Then for both cases, the gradient of $\boldsymbol{x}_{(i)}$ with respect to $\boldsymbol{w}_{(j)}$ can be bounded as follows:*

$$\left\|\frac{\partial \mathsf{vec}\left(\boldsymbol{X}_{(i)}\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\frac{\partial \boldsymbol{x}_{(i)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq d_i r_i c_i \tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{i}\frac{d_s {b_s}^2(k_s-s_s+1)^2}{16p^2}$$

and

$$\max_s\left\|\frac{\partial \mathsf{vec}\left(\boldsymbol{X}_{(i)}^s\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \max_s\left\|\frac{\partial \boldsymbol{x}_{(i)}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq r_i c_i \tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{i}\frac{d_s {b_s}^2(k_s-s_s+1)^2}{16p^2}.$$

Proof. For brevity, let $\boldsymbol{X}_{(i)}^k(s,t)$ denotes the (s,t)-th entry in the matrix $\boldsymbol{X}_{(i)}^k \in \mathbb{R}^{r_i\times c_i}$. We also let $\boldsymbol{\phi}_{(i,m)} = \frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{X}_{(m)}} \in \mathbb{R}^{r_m\times c_m\times d_m}$. So similar to Lemma 9, we have

$$\boldsymbol{\phi}_{(i,m)}^q = \mathsf{up}\left(\sum_{k=1}^{d_{m+1}}\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k\widetilde{\circledast}\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\right)\odot\sigma_1'(\boldsymbol{X}_{(m)}^q)\in\mathbb{R}^{r_m\times c_m},\quad (q=1,\cdots,d_m),$$

where the matrix $\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\in\mathbb{R}^{k_{m+1}\times k_{m+1}}$ is obtained by rotating $\boldsymbol{W}_{(m+1)}^{k,q}$ with 180 degrees. Then according to the relationship between $\boldsymbol{X}_{(j)}$ and $\boldsymbol{W}_{(j)}$, we can compute

$$\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{W}_{(j)}^{g,h}} = \boldsymbol{Z}_{(j-1)}^h\widetilde{\circledast}\boldsymbol{\phi}_{(i,j)}^g \in \mathbb{R}^{k_j\times k_j},\ (h=1,\cdots,d_{j-1};\ g=1,\cdots,d_j).$$

Therefore, we can further obtain

$$\begin{aligned}\left\|\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 =& \sum_{g=1}^{d_j}\sum_{h=1}^{d_{j-1}}\left\|\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{W}_{(j)}^{g,h}}\right\|_F^2 = \sum_{g=1}^{d_j}\sum_{h=1}^{d_{j-1}}\left\|\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{W}_{(j)}^{g,h}}\right\|_F^2 = \sum_{g=1}^{d_j}\sum_{h=1}^{d_{j-1}}\left\|\boldsymbol{Z}_{(j-1)}^h\widetilde{\circledast}\boldsymbol{\phi}_{(i,j)}^g\right\|_F^2 \\ \leq& \sum_{g=1}^{d_j}\sum_{h=1}^{d_{j-1}}(k_j-s_j+1)^2\left\|\boldsymbol{Z}_{(j-1)}^h\right\|_F^2\left\|\boldsymbol{\phi}_{(i,j)}^g\right\|_F^2 \leq (k_j-s_j+1)^2\left\|\boldsymbol{Z}_{(j-1)}\right\|_F^2\left\|\boldsymbol{\phi}_{(i,j)}\right\|_F^2 \\ \leq& \tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\left\|\boldsymbol{\phi}_{(i,j)}\right\|_F^2.\end{aligned}$$

On the other hand, by Lemma 9, we can further bound $\|\boldsymbol{\phi}_{(i,j)}\|_F^2$ as follows:

$$\begin{aligned}
\left\|\boldsymbol{\phi}_{(i,m)}\right\|_F^2 =& \sum_{q=1}^{d_m}\left\|\boldsymbol{\phi}_{(i,m)}^q\right\|_F^2 = \sum_{q=1}^{d_m}\left\|\mathsf{up}\left(\sum_{k=1}^{d_{m+1}}\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\right)\odot\sigma_1'(\boldsymbol{X}_{(m)}^q)\right\|_F^2 \\
\leq& \frac{1}{16p^2}\sum_{q=1}^{d_m}\left\|\sum_{k=1}^{d_{m+1}}\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k\widetilde{\circledast}\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\right\|_F^2 \leq \frac{d_{m+1}}{16p^2}\sum_{q=1}^{d_m}\sum_{k=1}^{d_{m+1}}\left\|\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k\widetilde{\circledast}\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\right\|_F^2 \\
\leq& \frac{d_{m+1}(k_{m+1}-s_{m+1}+1)^2}{16p^2}\sum_{q=1}^{d_m}\sum_{k=1}^{d_{m+1}}\left\|\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k\right\|_F^2\left\|\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\right\|_F^2 \\
\overset{①}{=}& \frac{d_{m+1}(k_{m+1}-s_{m+1}+1)^2}{16p^2}\sum_{k=1}^{d_{m+1}}\left\|\widetilde{\boldsymbol{\phi}}_{(i,m+1)}^k\right\|_F^2\left\|\widehat{\boldsymbol{W}}_{(m+1)}^{k}\right\|_F^2 \\
\leq& \frac{d_{m+1}{b_{m+1}}^2(k_{m+1}-s_{m+1}+1)^2}{16p^2}\left\|\boldsymbol{\phi}_{(i,m+1)}\right\|_F^2,
\end{aligned}$$

where ① holds since $\|\boldsymbol{W}_{(m+1)}^{k,q}\|_F^2 = \|\widehat{\boldsymbol{W}}_{(m+1)}^{k,q}\|_F^2$. It further yields

$$\left\|\boldsymbol{\phi}_{(i,m)}\right\|_F^2 \leq \left\|\boldsymbol{\phi}_{(i,i)}\right\|_F^2\prod_{s=m+1}^{i}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2} \overset{①}{=} \prod_{s=m+1}^{i}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}.$$

where ① holds since we have $\left\|\boldsymbol{\phi}_{(i,i)}\right\|_F^2 = \left\|\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{X}_{(i)}}\right\|_F^2 = 1$.

Therefore, we have

$$\begin{aligned}
\left\|\frac{\partial \boldsymbol{x}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 =& \sum_{s=1}^{r_i}\sum_{t=1}^{c_i}\left\|\frac{\partial \boldsymbol{X}_{(i)}^k(s,t)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \sum_{s=1}^{r_i}\sum_{t=1}^{c_i}\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\left\|\boldsymbol{\phi}_{(i,j)}\right\|_F^2 \\
=& r_ic_i\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\left\|\boldsymbol{\phi}_{(i,j)}\right\|_F^2 \\
\leq& r_ic_i\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{i}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}.
\end{aligned}$$

It further gives

$$\left\|\frac{\partial \boldsymbol{x}_{(i)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \sum_{s=1}^{d_i}\left\|\frac{\partial \boldsymbol{x}_{(i)}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq d_ir_ic_i\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{i}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}.$$

The proof is completed. □

Lemma 13. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w},\boldsymbol{D})$ is squared loss. Then the gradient of $\boldsymbol{\delta}_l^s$ with respect to $\boldsymbol{w}_{(j)}$ can be bounded as follows:*

$$\left\|\frac{\partial \boldsymbol{\delta}_l^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \frac{\tilde{\vartheta}{b_{l+1}}^2}{16p^2}\left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 d_lr_lc_l\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}$$

and

$$\left\|\frac{\partial \boldsymbol{\delta}_l}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \frac{\tilde{\vartheta}{b_{l+1}}^4}{16p^2}d_lr_lc_l\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2},$$

where $\tilde{\vartheta}=\frac{3}{64}$.

Proof. Assume that $\boldsymbol{W}_{(l+1)} = [\boldsymbol{W}_{(l+1)}^1, \boldsymbol{W}_{(l+1)}^2, \cdots, \boldsymbol{W}_{(l+1)}^{d_l}]$ where $\boldsymbol{W}_{(l+1)}^i \in \mathbb{R}^{d_{l+1} \times \tilde{r}_l \tilde{c}_l}$ is a submatrix in $\boldsymbol{W}_{(l+1)}$. Then we have $\boldsymbol{v} = \sigma_2(\sum_{k=1}^{d_l} \boldsymbol{W}_{(l+1)}^k \boldsymbol{z}_{(l)}^k)$. For brevity, we further define a matrix $\boldsymbol{G}_{(k)}$ as follows:

$$\boldsymbol{G}_{(k)} = \Big[\underbrace{\sigma_1'\left(\boldsymbol{x}_{(k)}\right), \sigma_1'\left(\boldsymbol{x}_{(k)}\right), \cdots, \sigma_1'\left(\boldsymbol{x}_{(k)}\right)}_{r_k c_k \text{ columns}}\Big] \in \mathbb{R}^{r_k c_k d_k \times r_k c_k},$$

Then we have

$$\frac{\partial}{\partial \boldsymbol{z}_{(l)}}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{z}_{(l)}^s}\right) = \left[(\boldsymbol{v}-\boldsymbol{y})^T \otimes (\boldsymbol{W}_{(l+1)}^s)^T\right] \mathsf{Q}(\boldsymbol{u}) \boldsymbol{W}_{(l+1)} + (\boldsymbol{W}_{(l+1)}^s)^T \mathsf{G}(\boldsymbol{u}) \mathsf{G}(\boldsymbol{u}) \boldsymbol{W}_{(l+1)},$$

where $\mathsf{Q}(\boldsymbol{u})$ is a matrix of size $d_{l+1}^2 \times d_{l+1}$ whose $(s, (s-1)d_{l+1}+s)$ entry equal to $\sigma_1(\boldsymbol{u}_s)(1-\sigma_1(\boldsymbol{u}_s))(1-2\sigma_1(\boldsymbol{u}_s))$ and rest entries are all 0. Accordingly, we have

$$\begin{aligned}
\left\|\frac{\partial}{\partial \boldsymbol{z}_{(l)}}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{z}_{(l)}^s}\right)\right\|_F^2 \leq & 2\left(\left\|\left[(\boldsymbol{v}-\boldsymbol{y})^T \otimes (\boldsymbol{W}_{(l+1)}^s)^T\right] \mathsf{Q}(\boldsymbol{u}) \boldsymbol{W}_{(l+1)}\right\|_F^2 + \left\|(\boldsymbol{W}_{(l+1)}^s)^T \mathsf{G}(\boldsymbol{u}) \mathsf{G}(\boldsymbol{u}) \boldsymbol{W}_{(l+1)}\right\|_F^2\right) \\
\leq & 2\left(\frac{2^6}{3^8}\|\boldsymbol{v}-\boldsymbol{y}\|_F^2 \left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 \left\|\boldsymbol{W}_{(l+1)}\right\|_F^2 + \frac{1}{16^2}\left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 \left\|\boldsymbol{W}_{(l+1)}\right\|_F^2\right) \\
\overset{①}{\leq} & 2\left(\frac{2^7}{3^8} + \frac{1}{16^2}\right) {b_{l+1}}^2 \left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 \\
\leq & \frac{3}{64} {b_{l+1}}^2 \left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2,
\end{aligned}$$

where we have $\|\boldsymbol{v}-\boldsymbol{y}\|_F^2 \leq 2$ by Lemma 8. Then by similar way, we can have

$$\begin{aligned}
\left\|\frac{\partial \boldsymbol{\delta}_l^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = & \left\|\frac{\partial}{\partial \boldsymbol{z}_{(l)}}\left(\frac{\partial f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{z}_{(l)}^s}\right) \frac{\partial \boldsymbol{z}_{(l)}}{\partial \boldsymbol{y}_l} \frac{\partial \boldsymbol{y}_l}{\partial \boldsymbol{x}_{(l)}} \frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \frac{3}{64 * 16p^2} {b_{l+1}}^2 \left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 \left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\leq & \frac{3}{64 * 16p^2} {b_{l+1}}^2 \left\|\boldsymbol{W}_{(l+1)}^s\right\|_F^2 d_l r_l c_l \tilde{r}_{j-1} \tilde{c}_{j-1} d_{j-1} (k_j - s_j + 1)^2 \prod_{s=j+1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2}.
\end{aligned}$$

Therefore, we can further obtain:

$$\left\|\frac{\partial \boldsymbol{\delta}_l}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \sum_{s=1}^{d_{l+1}} \left\|\frac{\partial \boldsymbol{\delta}_l^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \frac{3}{64 * 16p^2} {b_{l+1}}^4 d_l r_l c_l \tilde{r}_{j-1} \tilde{c}_{j-1} d_{j-1} (k_j - s_j + 1)^2 \prod_{s=j+1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2}.$$

The proof is completed. □

Lemma 14. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Then the Hessian of $f(\boldsymbol{w}, \boldsymbol{x})$ with respect to $\boldsymbol{w}$ can be bounded as follows:*

$$\left\|\nabla_{\boldsymbol{w}}^2 f(\boldsymbol{w}, \boldsymbol{D})\right\|_F^2 \leq \boldsymbol{\mathcal{O}}\left(\gamma^2\right),$$

where $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2}\right]^2\right)^{1/2}$. With the same condition, we can bound the operation norm of $\nabla_{\boldsymbol{w}}^3 f(\boldsymbol{w}, \boldsymbol{D})$. That is, there exists a universal constant ν such that $\left\|\nabla_{\boldsymbol{w}}^3 f(\boldsymbol{w}, \boldsymbol{D})\right\|_{op} \leq \nu$.

Proof. From Lemma 9, we can further compute the Hessian matrix $\nabla_{\boldsymbol{w}}^2 f(\boldsymbol{w}, \boldsymbol{D})$. Recall that $\boldsymbol{w}_{(i)}^k \in \mathbb{R}^{{k_i}^2 d_{i-1}}$ $(k = 1, \cdots, d_l)$ is the vectorization of $\boldsymbol{W}_{(i)}^k \in \mathbb{R}^{k_i \times k_i \times d_{i-1}}$, *i.e.* $\boldsymbol{w}_{(i)}^k = \left[\mathsf{vec}\left(\boldsymbol{W}_{(i)}^1(:,:1)\right); \cdots; \mathsf{vec}\left(\boldsymbol{W}_{(i)}^{d_i}(:,:,d_{i-1})\right)\right]$.

Let $\boldsymbol{w}_{(i)} = \left[\boldsymbol{w}_{(i)}^1; \cdots; \boldsymbol{w}_{(i)}^{d_i}\right] \in \mathbb{R}^{{k_i}^2 d_{i-1} \times d_i}$ $(i = 1, \cdots, l)$. Also, $\boldsymbol{w}_{(l+1)} \in \mathbb{R}^{\tilde{r}_l \tilde{c}_l d_l d_{l+1}}$ is the vectorization of the weight matrix $\boldsymbol{W}_{(l+1)}$. Then if $1 \leq i, j \leq l$, we can have

$$\begin{aligned}\frac{\partial^2 f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{w}_{(j)} \partial \boldsymbol{w}_{(i)}^k} &= \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{w}_{(j)} \partial \boldsymbol{w}_{(i)}^{k,1}} \\ \frac{\partial^2 f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{w}_{(j)} \partial \boldsymbol{w}_{(i)}^{k,2}} \\ \vdots \\ \frac{\partial^2 f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{w}_{(j)} \partial \boldsymbol{w}_{(i)}^{k,d_{i-1}}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^1 \widetilde{\circledast} \widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \\ \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^2 \widetilde{\circledast} \widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \\ \vdots \\ \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}} \widetilde{\circledast} \widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \end{bmatrix} \\ &= \begin{bmatrix} \mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right) \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^1\right)\right)}{\partial \boldsymbol{w}_{(j)}} + \mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^1\right) \frac{\partial \left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \\ \mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right) \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^2\right)\right)}{\partial \boldsymbol{w}_{(j)}} + \mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^2\right) \frac{\partial \left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \\ \vdots \\ \mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right) \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right)\right)}{\partial \boldsymbol{w}_{(j)}} + \mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right) \frac{\partial \left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \end{bmatrix} \in \mathbb{R}^{{k_i}^2 d_{i-1} \times {k_j}^2 d_j d_{j-1}},\end{aligned} \tag{4}$$

where $\mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right) \in \mathbb{R}^{{k_i}^2 \times \tilde{r}_{i-1} \tilde{c}_{i-1} d_{i-1}}$ and $\mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right) \in \mathbb{R}^{{k_i}^2 \times (\tilde{r}_i - k_i + 1)(\tilde{c}_i - k_i + 1)}$ satisfy: each row in $\mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right)$ contains the vectorization of $(\widetilde{\boldsymbol{\delta}}_i^k)^T$ at the right position and the remaining entries are 0s, and each row in $\mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right)$ is the submatrix in $\boldsymbol{Z}_{(i-1)}^{d_{i-1}}$ that need to conduct inner product with $\widetilde{\boldsymbol{\delta}}_i^k$ in turn. Note that there are $s_i - 1$ rows and columns between each neighboring nonzero entries in $\boldsymbol{N}$ which is decided by the definition of $\widetilde{\boldsymbol{\delta}}_{i+1}$ in Sec. B.1. Accordingly, we have

$$\left\| \mathsf{P}_1\left(\widetilde{\boldsymbol{\delta}}_i^k\right) \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2 \leq (k_i - s_i + 1)^2 \|\widetilde{\boldsymbol{\delta}}_i^k\|_F^2 \left\| \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2 = (k_i - s_i + 1)^2 \|\boldsymbol{\delta}_i^k\|_F^2 \left\| \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2$$

and

$$\left\| \mathsf{P}_2\left(\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\right) \frac{\partial \left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2 \leq (k_i - s_i + 1)^2 \|\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\|_F^2 \left\| \frac{\partial \left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2 = (k_i - s_i + 1)^2 \|\boldsymbol{Z}_{(i-1)}^{d_{i-1}}\|_F^2 \left\| \frac{\partial \left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}} \right\|_F^2.$$

Then in order to bound

$$\|\nabla_{\boldsymbol{w}}^2 f(\boldsymbol{w}, \boldsymbol{D})\|_F^2 = \sum_{i=1}^{l+1} \sum_{j=1}^{l+1} \left\| \frac{\partial^2 f(\boldsymbol{w}, \boldsymbol{D})}{\partial \boldsymbol{w}_{(j)} \partial \boldsymbol{w}_{(i)}} \right\|_F^2,$$

we try to bound each term separately. So we consider the following five cases: $l \geq i \geq j$, $i \leq j \leq l$, $l + 1 = i > j$, $l + 1 = j > i$ and $l + 1 = i = j$.

Case 1: $l \geq i \geq j$

In the following, we first consider the first case, *i.e.* $i \geq j$, and bound

$$\begin{aligned}
&\left\|\frac{\partial\left(\mathsf{vec}\left(\widetilde{\boldsymbol{\delta}}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\frac{\partial}{\partial \boldsymbol{w}_{(j)}}\mathsf{vec}\left[\mathsf{up}\left(\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right)\odot\sigma_1'(\boldsymbol{X}_{(i)}^k)\right]\right\|_F^2 \\
\overset{①}{\leq}&\frac{2}{16}\left\|\frac{\partial}{\partial \boldsymbol{w}_{(j)}}\mathsf{vec}\left[\mathsf{up}\left(\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right)\right]\right\|_F^2 + 2\left\|\mathsf{up}\left(\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right)\right\|_F^2\left\|\frac{\partial\sigma_1'(\boldsymbol{X}_{(i)}^k)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
=&\frac{2}{16p^2}\left\|\frac{\partial}{\partial \boldsymbol{w}_{(j)}}\mathsf{vec}\left[\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right]\right\|_F^2 + \frac{2}{p^2}\left\|\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right\|_F^2\left\|\frac{\partial\mathsf{vec}\left(\sigma'(\boldsymbol{x}_{(i)}^k)\right)}{\partial \boldsymbol{x}_{(i)}^k}\frac{\partial \boldsymbol{x}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\overset{②}{\leq}&\frac{2}{16p^2}\left\|\frac{\partial}{\partial \boldsymbol{w}_{(j)}}\mathsf{vec}\left[\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right]\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}\left\|\sum_{s=1}^{d_{i+1}}\widetilde{\boldsymbol{\delta}}_{i+1}^s \widetilde{\circledast}\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right\|_F^2\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\overset{③}{\leq}&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2\sum_{s=1}^{d_{i+1}}\|\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\|_F^2\left\|\frac{\partial\widetilde{\boldsymbol{\delta}}_{i+1}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}d_{i+1}(k_{i+1}-s_{i+1}+1)^2\sum_{s=1}^{d_{i+1}}\left\|\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}\right\|_F^2\left\|\widetilde{\boldsymbol{\delta}}_{i+1}^s\right\|_F^2\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\overset{④}{=}&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2\sum_{s=1}^{d_{i+1}}\|\boldsymbol{W}_{(i+1)}^{s,k}\|_F^2\left\|\frac{\partial\boldsymbol{\delta}_{i+1}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}d_{i+1}(k_{i+1}-s_{i+1}+1)^2\sum_{s=1}^{d_{i+1}}\left\|\boldsymbol{W}_{(i+1)}^{s,k}\right\|_F^2\left\|\boldsymbol{\delta}_{i+1}^s\right\|_F^2\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2.
\end{aligned} \tag{5}$$

① holds since $\boldsymbol{X}_{(i)}^k$ is independent on $\boldsymbol{w}_{(j)}$ and the values of entries in $\sigma_1'(\boldsymbol{X}_{(i)}^k)$ is not larger than $1/4$ since for any constant a, $\sigma'(a)=\sigma(a)(1-\sigma(a))\leq 1/4$. ② holds since for arbitrary tensor $\boldsymbol{M}$, we have $\|\mathsf{up}(\boldsymbol{M})\|_F^2 \leq \|\boldsymbol{M}\|_F^2/p^2$ in Lemma 8, and we also have

$$\left\|\frac{\partial\mathsf{vec}\left(\sigma'(\boldsymbol{x}_{(i)}^k)\right)}{\partial \boldsymbol{x}_{(i)}^k}\frac{\partial \boldsymbol{x}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\mathsf{Q}\left(\boldsymbol{x}_{(i)}^k\right)\frac{\partial \boldsymbol{x}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \leq \frac{2^6}{3^8}\left\|\frac{\partial \boldsymbol{x}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \frac{2^6}{3^8}\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2.$$

③ holds since we can just adopt similar strategy in Eqn. (4) to separate $\widehat{\boldsymbol{W}}_{(i+1)}^{s,k}$ and the conclusion in Lemma 8; ④ holds since the difference between $\widetilde{\boldsymbol{\delta}}_{i+1}^s$ and $\boldsymbol{\delta}_{i+1}^s$ is that we pad 0 around $\boldsymbol{\delta}_{i+1}^s$ to obtain $\widetilde{\boldsymbol{\delta}}_{i+1}^s$, indicating $\|\boldsymbol{\delta}_{i+1}^s\|_F^2 = \|\widetilde{\boldsymbol{\delta}}_{i+1}^s\|_F^2$.

Accordingly, we can further bound

$$\begin{aligned}
&\left\|\frac{\partial\widetilde{\boldsymbol{\delta}}_i}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \sum_{k=1}^{d_{i-1}}\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\leq&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2\sum_{k=1}^{d_{i-1}}\sum_{s=1}^{d_{i+1}}\|\boldsymbol{W}_{(i+1)}^{s,k}\|_F^2\left\|\frac{\partial\boldsymbol{\delta}_{i+1}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}d_{i+1}(k_{i+1}-s_{i+1}+1)^2\sum_{k=1}^{d_{i-1}}\sum_{s=1}^{d_{i+1}}\left\|\boldsymbol{W}_{(i+1)}^{s,k}\right\|_F^2\left\|\boldsymbol{\delta}_{i+1}^s\right\|_F^2\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\leq&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2\sum_{s=1}^{d_{i+1}}\|\boldsymbol{W}_{(i+1)}^{s}\|_F^2\left\|\frac{\partial\boldsymbol{\delta}_{i+1}^s}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}d_{i+1}(k_{i+1}-s_{i+1}+1)^2\left\|\boldsymbol{\delta}_{i+1}\right\|_F^2\max_s\left\|\boldsymbol{W}_{(i+1)}^{s}\right\|_F^2\max_k\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\overset{①}{\leq}&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2{b_{i+1}}^2\left\|\frac{\partial\boldsymbol{\delta}_{i+1}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{2\cdot 2^6}{3^8p^2}d_{i+1}(k_{i+1}-s_{i+1}+1)^2{b_{i+1}}^2\left\|\boldsymbol{\delta}_{i+1}\right\|_F^2\max_k\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\overset{②}{\leq}&\frac{2d_{i+1}}{16p^2}(k_{i+1}-s_{i+1}+1)^2{b_{i+1}}^2\left\|\frac{\partial\boldsymbol{\delta}_{i+1}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 + \frac{\vartheta{b_{l+1}}^2 d_{j-1}}{3p^2{b_j}^2 d_j}r_i c_i r_{j-1}c_{j-1}\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2},
\end{aligned}$$

where ① holds since we have $\|\boldsymbol{W}_{(i+1)}^s\|_F \leq r_w$; ② holds due to the bounds of $\left\|\boldsymbol{\delta}_{i+1}\right\|_F^2$ and $\left\|\frac{\partial \boldsymbol{X}_{(i)}^k}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2$ in Lemma 10 and 12.

Then, we can use the above recursion inequality to further obtain

$$
\begin{aligned}
&\left\|\frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\
\leq&\left[\prod_{s=i+1}^{i+1}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]\left(\frac{2d_{i+2}}{16p^2}(k_{i+2}-s_{i+2}+1)^2{b_{i+2}}^2\left\|\frac{\partial \boldsymbol{\delta}_{i+2}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2+\frac{\vartheta{b_{l+1}}^2d_{j-1}}{3p^2{b_j}^2d_j}r_{j-1}c_{j-1}r_{i+1}c_{i+1}\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\right)\\
&+\frac{\vartheta{b_{l+1}}^2d_{j-1}}{3p^2{b_j}^2d_j}r_ic_ir_{j-1}c_{j-1}\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\\
\leq&\left[\prod_{s=i+1}^{l}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]\left\|\frac{\partial \boldsymbol{\delta}_l}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2+\frac{\vartheta{b_{l+1}}^2d_{j-1}}{3p^2{b_j}^2d_j}r_{j-1}c_{j-1}\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\left[r_ic_i+r_{i+1}c_{i+1}\left[\prod_{s=i+1}^{i+1}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]\right.\\
&\left.+r_{i+2}c_{i+2}\left[\prod_{s=i+1}^{i+2}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]+\cdots+r_lc_l\left[\prod_{s=i+1}^{l}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]\right].
\end{aligned}
$$

By Lemma 13, we have

$$
\left\|\frac{\partial \boldsymbol{\delta}_l}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\leq\frac{\tilde{\vartheta}{b_{l+1}}^4d_{j-1}}{p^2{b_j}^2d_j}d_lr_lc_lr_{j-1}c_{j-1}\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2},
$$

where $\tilde{\vartheta}=\frac{3}{64}$. Thus, we can establish

$$
\left\|\frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\leq\frac{\tilde{\vartheta}{b_{l+1}}^2d_{j-1}}{p^2{b_j}^2d_j}r_{j-1}c_{j-1}\left[\frac{\tau}{3}+{b_{l+1}}^2d_lr_lc_l\prod_{s=i+1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\right]\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}.
$$

where $\tau=r_ic_i+r_{i+1}c_{i+1}\left[\prod_{s=i+1}^{i+1}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]+\cdots+r_lc_l\left[\prod_{s=i+1}^{l}\frac{2d_s}{16p^2}(k_s-s_s+1)^2{b_s}^2\right]$. It further gives the bound of $\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{x})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(i)}}\right\|_F^2$ as follows:

$$
\begin{aligned}
\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(i)}}\right\|_F^2=&\sum_{k=1}^{d_{i-1}}\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(i)}^k}\right\|_F^2=\sum_{k=1}^{d_{i-1}}\sum_{s=1}^{d_i}\left\|\mathsf{P}_1\left(\boldsymbol{\delta}_i^k\right)\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{X}_{(i-1)}^s\right)\right)}{\partial \boldsymbol{w}_{(j)}}+\mathsf{P}_2\left(\boldsymbol{X}_{(i-1)}^s\right)\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\\
\leq&2\sum_{k=1}^{d_{i-1}}\sum_{s=1}^{d_i}\left(\left\|\mathsf{P}_1\left(\boldsymbol{\delta}_i^k\right)\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{X}_{(i-1)}^s\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2+\left\|\mathsf{P}_2\left(\boldsymbol{X}_{(i-1)}^s\right)\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\right)\\
\leq&2(k_i-s_i+1)^2\sum_{k=1}^{d_{i-1}}\sum_{s=1}^{d_i}\left(\left\|\boldsymbol{\delta}_i^k\right\|_F^2\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{X}_{(i-1)}^s\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2+\left\|\boldsymbol{X}_{(i-1)}^s\right\|_F^2\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i^k\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\right)\\
\leq&2(k_i-s_i+1)^2\left(\left\|\boldsymbol{\delta}_i\right\|_F^2\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{X}_{(i-1)}\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2+\left\|\boldsymbol{X}_{(i-1)}\right\|_F^2\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\right)\\
\overset{①}{\leq}&\frac{32\vartheta{b_{l+1}}^2d_{i-1}}{{b_i}^2{b_j}^2d_id_j}r_{i-1}c_{i-1}r_{j-1}c_{j-1}\prod_{s=i+1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}+r_{i-1}c_{i-1}d_{i-1}\left\|\frac{\partial\left(\mathsf{vec}\left(\boldsymbol{\delta}_i\right)\right)}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\\
\overset{②}{\leq}&\boldsymbol{\mathcal{O}}\left(\frac{\vartheta{b_{l+1}}^2d_{i-1}d_{j-1}}{{b_i}^2{b_j}^2d_id_j}r_{i-1}c_{i-1}r_{j-1}c_{j-1}\left[\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\right]\left[\prod_{s=i}^{l}\frac{2d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\right]\right).
\end{aligned}
$$

where ① holds because of Lemma 12, while ② holds due to 13.

Case 2: $i \leq j \leq l$

Since $\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^T}$ is symmetrical, we have $\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(i)}^T \partial \boldsymbol{w}_{(j)}} = \left(\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}^T \partial \boldsymbol{w}_{(i)}}\right)^T$ $(1 \leq i,j \leq l)$. Thus, it yields

$$\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(i)}^T \partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}^T \partial \boldsymbol{w}_{(i)}}\right\|_F^2.$$

Case 3: $l+1 = i > j$

In the following, we first consider the first case, *i.e.* cross entropy and softmax activation, and bound

$$\begin{aligned}\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2 &= \left\|\frac{\partial(\boldsymbol{v}-\boldsymbol{y})\boldsymbol{z}_{(l)}^T}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 = \left\|\left[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\frac{\partial \boldsymbol{z}_{(l)}^T}{\partial \boldsymbol{x}_{(l)}}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}} + \left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{z}_{(l)}}\frac{\partial \boldsymbol{z}_{(l)}}{\partial \boldsymbol{x}_{(l)}}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &= \left\|\widetilde{\mathsf{up}}\left(\left[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right] + \left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\mathsf{diag}\left(\sigma_2'(\boldsymbol{u})\right)\boldsymbol{W}_{(l+1)}\right)\odot \boldsymbol{G}_{(l)}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2,\end{aligned}$$

where $\boldsymbol{G}_{(l)}$ is defined as

$$\boldsymbol{G}_{(l)} = \Big[\underbrace{\sigma_1'\left(\boldsymbol{x}_{(l)}\right),\sigma_1'\left(\boldsymbol{x}_{(l)}\right),\cdots,\sigma_1'\left(\boldsymbol{x}_{(l)}\right)}_{r_l c_l \text{ columns}}\Big] \in \mathbb{R}^{r_l c_l d_l \times r_l c_l}.$$

Thus, we can further obtain

$$\begin{aligned}\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2 &\leq \frac{1}{16p^2}\left\|\left[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right] + \left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\mathsf{diag}\left(\sigma_2'(\boldsymbol{u})\right)\boldsymbol{W}_{(l+1)}\right\|_F^2\left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &\leq \frac{2}{16p^2}\left(\left\|\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right\|_F^2 + \left\|\left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\mathsf{diag}\left(\sigma_2'(\boldsymbol{u})\right)\boldsymbol{W}_{(l+1)}\right\|_F^2\right)\left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &\overset{①}{\leq} \frac{2}{16p^2}\left(\left\|\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right\|_F^2 + \left\|\boldsymbol{z}_{(l)}\otimes\left[\mathsf{diag}\left(\sigma_2'(\boldsymbol{u})\right)\boldsymbol{W}_{(l+1)}\right]\right\|_F^2\right)\left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &\overset{②}{\leq} \frac{1}{8p^2}\tilde{r}_l\tilde{c}_l d_l\left(2+\frac{1}{16}{b_{l+1}}^2\right)d_l r_l c_l \tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{l}\frac{d_s {b_s}^2(k_s-s_s+1)^2}{16p^2} \\ &= \frac{2d_{j-1}}{p^4 {b_j}^2 d_j} r_l^2 c_l^2 d_l^2 r_{j-1}c_{j-1}\left(2+\frac{1}{16}{b_{l+1}}^2\right)\prod_{s=j}^{l}\frac{d_s {b_s}^2(k_s-s_s+1)^2}{16p^2}\end{aligned}$$

where ① holds since for an arbitrary vector $\boldsymbol{u}\in\mathbb{R}^k$ and an arbitrary matrix $\boldsymbol{M}\in\mathbb{R}^{k\times k}$, we have $(\boldsymbol{u}\otimes \boldsymbol{I}_k)\boldsymbol{M} = \boldsymbol{u}\otimes\boldsymbol{M}$; ② holds since we use Lemma 12 and the assumption that $\|\boldsymbol{W}_{(l+1)}\|_F^2 \leq {b_{l+1}}^2$.

Now we consider the least square loss and softmax activation function. In such a case, we can further obtain:

$$\begin{aligned}&\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2 = \left\|\frac{\partial(\boldsymbol{v}-\boldsymbol{y})\mathsf{G}(\boldsymbol{u})\boldsymbol{z}_{(l)}^T}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &= \left\|\left[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\frac{\partial \boldsymbol{z}_{(l)}^T}{\partial \boldsymbol{x}_{(l)}}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}} + \left[\boldsymbol{z}_{(l)}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\frac{\partial \mathsf{vec}\left(\mathsf{G}(\boldsymbol{u})\right)}{\partial \boldsymbol{u}}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{z}_{(l)}}\frac{\partial \boldsymbol{z}_{(l)}}{\partial \boldsymbol{x}_{(l)}}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}} + \left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{z}_{(l)}}\frac{\partial \boldsymbol{z}_{(l)}}{\partial \boldsymbol{x}_{(l)}}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2 \\ &= \left\|\widetilde{\mathsf{up}}\left(\left[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\right] + \left[\boldsymbol{z}_{(l)}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\mathsf{Q}(\boldsymbol{u})\boldsymbol{W}_{(l+1)} + \left[\boldsymbol{z}_{(l)}\otimes \boldsymbol{I}_{d_{l+1}}\right]\mathsf{Q}(\boldsymbol{u})\mathsf{G}(\boldsymbol{u})\boldsymbol{W}_{(l+1)}\right)\odot \boldsymbol{G}_{(l)}\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2.\end{aligned}$$

Thus, we can further obtain

$$
\begin{aligned}
&\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2\\
\leq&\frac{1}{16p^2}\left\|[\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})]+\left[\boldsymbol{z}_{(l)}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\mathsf{Q}(\boldsymbol{u})\boldsymbol{W}_{(l+1)}+\left[\boldsymbol{z}_{(l)}\otimes\boldsymbol{I}_{d_{l+1}}\right]\mathsf{Q}(\boldsymbol{u})\mathsf{G}(\boldsymbol{u})\boldsymbol{W}_{(l+1)}\right\|_F^2\left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\\
\leq&\frac{3}{16p^2}\left(\|\boldsymbol{I}_{\tilde{r}_l\tilde{c}_l d_l}\otimes(\boldsymbol{v}-\boldsymbol{y})\|_F^2+\left\|\left[\boldsymbol{z}_{(l)}\otimes(\boldsymbol{v}-\boldsymbol{y})\right]\mathsf{Q}(\boldsymbol{u})\boldsymbol{W}_{(l+1)}\right\|_F^2+\left\|\left[\boldsymbol{z}_{(l)}\otimes\boldsymbol{I}_{d_{l+1}}\right]\mathsf{Q}(\boldsymbol{u})\mathsf{G}(\boldsymbol{u})\boldsymbol{W}_{(l+1)}\right\|_F^2\right)\left\|\frac{\partial \boldsymbol{x}_{(l)}}{\partial \boldsymbol{w}_{(j)}}\right\|_F^2\\
\overset{①}{\leq}&\frac{3}{16p^2}\tilde{r}_l\tilde{c}_l d_l\left(2+\frac{3}{100}{b_{l+1}}^2\right)d_l r_l c_l\tilde{r}_{j-1}\tilde{c}_{j-1}d_{j-1}(k_j-s_j+1)^2\prod_{s=j+1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\\
=&\frac{3d_{j-1}}{p^4{b_j}^2 d_j}r_{j-1}c_{j-1}d_l^2 r_l^2 c_l^2\left(2+\frac{2^7}{3^8}{b_{l+1}}^2+\frac{2^6}{16\cdot 3^8}{b_{l+1}}^2\right)\prod_{s=j}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2},
\end{aligned}
$$

where ① holds since we use Lemma 12 and the fact that $\|\boldsymbol{W}_{(l+1)}\|_F^2\leq {b_{l+1}}^2$.

Case 4: $i<j=l+1$

Similar to the Case 2, we also can have

$$
\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(i)}^T\partial \boldsymbol{w}_{(j)}}\right\|_F^2=\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}^T\partial \boldsymbol{w}_{(i)}}\right\|_F^2.
$$

So in this case, we can just directly use the bound in case 3 to bound $\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(i)}^T\partial \boldsymbol{w}_{(j)}}\right\|_F^2$.

Case 5: $i=j=l+1$

In the following, we first consider the first case, *i.e.* $i=l+1$, and bound

$$
\begin{aligned}
\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(l+1)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2=&\left\|\frac{\partial(\boldsymbol{v}-\boldsymbol{y})\boldsymbol{z}_{(l)}^T}{\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2=\left\|\left[\boldsymbol{z}_{(l)}\otimes\boldsymbol{I}_{d_{l+1}}\right]\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2\\
=&\left\|\left[\boldsymbol{z}_{(l)}\otimes\boldsymbol{I}_{d_{l+1}}\right]\mathsf{G}(\boldsymbol{u})\left[\boldsymbol{z}_{(l)}\otimes\boldsymbol{I}_{d_{l+1}}\right]^T\right\|_F^2\\
\overset{①}{=}&\left\|\left[\boldsymbol{z}_{(l)}\left(\boldsymbol{z}_{(l)}\otimes\mathsf{G}(\boldsymbol{u})\right)^T\right]^T\right\|_F^2\\
\leq&\left\|\boldsymbol{z}_{(l)}\right\|_F^4\|\mathsf{G}(\boldsymbol{u})\|_F^2\leq\frac{1}{16}\tilde{r}_l^2\tilde{c}_l^2 d_l^2 d_{l+1},
\end{aligned}
$$

where ① holds since for an arbitrary vector $\boldsymbol{u}\in\mathbb{R}^k$ and an arbitrary matrix $\boldsymbol{M}\in\mathbb{R}^{k\times k}$, we have $(\boldsymbol{u}\otimes\boldsymbol{I}_k)\boldsymbol{M}=\boldsymbol{u}\otimes\boldsymbol{M}$.

Now we can bound $\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}\partial \boldsymbol{w}}\right\|_F^2$ as follows:

$$
\begin{aligned}
\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}\partial \boldsymbol{w}}\right\|_F^2=&\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(l+1)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2+2\sum_{j=1}^{l}\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(l+1)}}\right\|_F^2+2\sum_{j=1}^{l}\sum_{i=j}^{l}\left\|\frac{\partial^2 f(\boldsymbol{w},\boldsymbol{D})}{\partial \boldsymbol{w}_{(j)}\partial \boldsymbol{w}_{(i)}}\right\|_F^2\\
\leq&\boldsymbol{\mathcal{O}}\left(\frac{l^2}{{k_1}^4}\max_{1\leq i,j\leq l}\tilde{r}_i\tilde{c}_i r_j c_j\frac{{b_{l+1}}^4}{16p^2}\left(\frac{{r_w}^2}{8\sqrt{2}p^2}\right)^{2l-2}\prod_{s=1}^{l}({d_s{k_s}^2})^2\right)\\
\leq&\boldsymbol{\mathcal{O}}\left(\frac{\vartheta{b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2}l^2 r_0^2 c_0^2\left[\prod_{s=1}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{8\sqrt{2}p^2}\right]^2\right).
\end{aligned}
$$

On the other hand, if the activation functions σ_1 and σ_2 are respectively sigmoid function and softmax function, $f(\boldsymbol{w}, \boldsymbol{D})$ is infinitely differentiable. Also $\sigma(a)$, $\sigma'(a)$, $\sigma''(a)$ and $\sigma'''(a)$ are all bounded. This means that $\nabla^3_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D})$ exists. Also since input $\boldsymbol{D}$ and the parameter $\boldsymbol{w}$ are bounded, we can always find a universal constant ν such that

$$\|\nabla^3_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D})\|_{\text{op}} = \sup_{\|\boldsymbol{\lambda}\|_2 \leq 1} \Big\langle \boldsymbol{\lambda}^{\otimes^3}, \nabla^3_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}) \Big\rangle = \sum_{i,j,k} [\nabla^3_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D})]_{ijk} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_j \boldsymbol{\lambda}_k \leq \nu < +\infty.$$

The proof is completed. □

Lemma 15. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Suppose Assumption 1 on the input data $\boldsymbol{D}$ holds. Then for any $t > 0$, the objective $f(\boldsymbol{w}, \boldsymbol{x})$ obeys*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \Big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}(f(\boldsymbol{w}, \boldsymbol{D}^{(i)})) \Big) > t \right) \leq 2 \exp\left(-\frac{2nt^2}{\alpha^2} \right),$$

where $\alpha = 1$.

Proof. Since the input $\boldsymbol{D}^{(i)}$ $(i = 1, \cdots, n)$ are independent from each other, then the output $f(\boldsymbol{w}, \boldsymbol{D}^{(i)})$ $(i = 1, \cdots, n)$ are also independent. Meanwhile, when the loss is the square loss, we can easily bound $0 \leq f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{y}\|_2^2 \leq 1$, since the value of entries in $\boldsymbol{v}$ belongs to $[0, 1]$ and $\boldsymbol{y}$ is a one-hot vector label of $\boldsymbol{v}$.

Besides, for arbitrary random variable x, $|x - \mathbb{E}x| \leq |x|$. So by Hoeffding's inequality in Lemma 6, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \Big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}(f(\boldsymbol{w}, \boldsymbol{D}^{(i)})) \Big) > t \right) \leq \exp\left(-\frac{2nt^2}{\alpha^2} \right),$$

where $\alpha = 1$. This means that $\frac{1}{n}\sum_{i=1}^{n} \big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}(f(\boldsymbol{w}, \boldsymbol{D}^{(i)})) \big)$ has exponential tails. □

Lemma 16. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Suppose Assumption 1 on the input data $\boldsymbol{D}$ holds. Then for any $t > 0$ and arbitrary unit vector $\boldsymbol{\lambda} \in \mathbb{S}^{d-1}$, the gradient $\nabla f(\boldsymbol{w}, \boldsymbol{x})$ obeys*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \Big(\Big\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) \Big\rangle \Big) > t \right) \leq \exp\left(-\frac{nt^2}{2\beta^2} \right).$$

where $\beta \triangleq \Big[\vartheta \tilde{r}_l \tilde{c}_l d_l + \sum_{i=1}^{l} \frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1} c_{i-1} \prod_{s=i}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16 p^2} \Big]^{1/2}$ in which $\vartheta = 1/8$.

Proof. Since the input $\boldsymbol{D}^{(i)}$ $(i = 1, \cdots, n)$ are independent from each other, then the output $\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})$ $(i = 1, \cdots, n)$ are also independent. Furthermore, for arbitrary vector $\boldsymbol{x}$, $\|\boldsymbol{x} - \mathbb{E}\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\|_2^2$. Hence, for an arbitrary unit vector $\boldsymbol{\lambda} \in \mathbb{S}^{d-1}$ where $d = \tilde{r}_l \tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, we have

$$\begin{aligned} \langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) \rangle \leq & \|\boldsymbol{\lambda}\|_2 \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\|_2 \\ \leq & \|\boldsymbol{\lambda}\|_2 \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\|_2 \overset{①}{\leq} \beta, \end{aligned}$$

where ① holds since $\|\boldsymbol{\lambda}\|_2 = 1$ $(\boldsymbol{\lambda} \in \mathbb{S}^{d-1})$ and by Lemma 11, we have $\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\| \leq \beta$ where $\beta \triangleq \Big[\vartheta \tilde{r}_l \tilde{c}_l d_l + \sum_{i=1}^{l} \frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1} c_{i-1} \prod_{s=i}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16 p^2} \Big]^{1/2}$ in which $\vartheta = 1/8$.

Thus, we can use Hoeffding's inequality in Lemma 6 to bound

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \Big(\Big\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) \Big\rangle \Big) > t \right) \leq \exp\left(-\frac{nt^2}{2\beta^2} \right).$$

The proof is completed. □

Lemma 17. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Suppose that Assumption 1 on the input data $\boldsymbol{D}$ and the parameter $\boldsymbol{w}$ holds. Then for any $t > 0$ and arbitrary unit vector $\boldsymbol{\lambda} \in \mathbb{S}^{d-1}$, the Hessian $\nabla^2 f(\boldsymbol{w}, \boldsymbol{D})$ obeys*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle \boldsymbol{\lambda}, (\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}))\boldsymbol{\lambda}\right\rangle\right) > t\right) \leq 2\exp\left(-\frac{nt^2}{2\gamma^2}\right).$$

where $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2}\right]^2\right)^{1/2}$.

Proof. Since the input $\boldsymbol{D}^{(i)}$ $(i = 1, \cdots, n)$ are independent from each other, then the output $\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})$ $(i = 1, \cdots, n)$ are also independent. On the other hand, for arbitrary random matrix $\boldsymbol{X}$, $\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|_F^2 \leq \|\boldsymbol{X}\|_F^2$. Thus, for an arbitrary unit vector $\boldsymbol{\lambda} \in \mathbb{S}^{d-1}$ where $d = \tilde{r}_l \tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, we have

$$\begin{aligned}
\langle \boldsymbol{\lambda}, (\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}))\boldsymbol{\lambda}\rangle \leq & \|\boldsymbol{\lambda}\|_2 \|(\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}))\boldsymbol{\lambda}\|_2 \\
\leq & \|\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\|_{\text{op}} \|\boldsymbol{\lambda}\|_2 \\
\leq & \|\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\|_F \\
\leq & \|\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\|_F \\
\overset{①}{\leq} & \gamma,
\end{aligned}$$

where ① holds since $\|\boldsymbol{\lambda}\|_2 = 1$ $(\boldsymbol{\lambda} \in \mathbb{S}^{d-1})$ and by Lemma 14, we have $\|\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\| \leq \gamma$ where $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2}\right]^2\right)^{1/2}$.

Thus, we can use Hoeffding's inequality in Lemma 6 to bound

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle \boldsymbol{\lambda}, (\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}))\boldsymbol{\lambda}\right\rangle\right) > t\right) \leq \exp\left(-\frac{nt^2}{2\gamma^2}\right).$$

The proof is completed. □

Lemma 18. *Suppose that the activation function σ_1 is sigmoid and σ_2 is softmax, and the loss function $f(\boldsymbol{w}, \boldsymbol{D})$ is squared loss. Suppose that Assumption 1 on the input data $\boldsymbol{D}$ and the parameter $\boldsymbol{w}$ holds. Then the empirical Hessian converges uniformly to the population Hessian in operator norm. Specifically, there exit two universal constants $c_{v'}$ and c_v such that if $n \geq c_{v'} \frac{\nu^2}{d\varrho\varepsilon^2}\left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2}\right]^{-1}$, then with probability at least $1 - \epsilon$*

$$\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla^2 \boldsymbol{Q}(\boldsymbol{w})\right\|_{op} \leq c_v \gamma \sqrt{\frac{2d + \theta\varrho + \log\left(\frac{4}{\varepsilon}\right)}{2n}},$$

holds with probability at least $1 - \varepsilon$, where $d = \tilde{r}_l \tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, $\theta = a_{l+1}(d_{l+1} + \tilde{r}_l \tilde{c}_l d_l - 2a_{l+1} + 1) + \sum_{i=1}^{l} a_i({k_i}^2 d_i + d_{i-1} - 2a_i + 1)$, $\varrho = \sum_{i=1}^{l} \log\left(\frac{\sqrt{d_i} b_i (k_i - s_i + 1)}{4p}\right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2}\right)$, and $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2}\right]^2\right)^{1/2}$.

Proof. Recall that the weight of each kernel and the feature maps has magnitude bound separately, *i.e.* $\boldsymbol{w}_{(i)}^k \in \mathsf{B}^{{k_i}^2 d_{i-1}}(r_w)$ $(i = 1, \cdots, l; k = 1, \cdots, d_i)$ and $\boldsymbol{w}_{(l+1)} \in \mathsf{B}^{\tilde{r}_l \tilde{c}_l d_l d_{l+1}}(b_{l+1})$. Since $\widetilde{\boldsymbol{W}}_{(i)} = [\mathsf{vec}(\boldsymbol{W}_{(i)}^1), \mathsf{vec}(\boldsymbol{W}_{(i)}^2), \cdots, \mathsf{vec}(\boldsymbol{W}_{(i)}^{d_{i-1}})] \in \mathbb{R}^{k_i^2 d_i \times d_{i-1}}$, we have $\|\widetilde{\boldsymbol{W}}_{(i)}\|_F \leq d_i b_i$.

So here we assume $\widetilde{\boldsymbol{W}}_{(i,\epsilon)}$ is the $d_i b_i \epsilon/(b_{l+1}+\sum_{i=1}^{l} d_i b_i)$-covering net of the matrix $\widetilde{\boldsymbol{W}}_{(i)}$ which is the set of all parameters in the i-th layer. Then by Lemma 7, we have the covering number

$$n_\epsilon{}^i \leq \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_i(k_i{}^2 d_i + d_{i-1} - 2a_i + 1)},$$

since the rank of $\widetilde{\boldsymbol{W}}_{(i)}$ obeys $\mathsf{rank}(\widetilde{\boldsymbol{W}}_{(i)}) \leq a_i$ for $1 \leq i \leq l$. For the last layer, we also can construct an $b_{l+1}\epsilon/(b_{l+1}+\sum_{i=1}^{l} d_i b_i)$-covering net for the weight matrix $\boldsymbol{W}_{(l+1)}$. Here we have

$$n_\epsilon{}^{l+1} \leq \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1)},$$

since the rank of $\boldsymbol{W}_{(l+1)}$ obeys $\mathsf{rank}(\boldsymbol{W}_{(l+1)}) \leq a_{l+1}$. Finally, we arrange them together to construct a set Θ and claim that there is always an ϵ-covering net $\boldsymbol{w}_\epsilon$ in Θ for any parameter $\boldsymbol{w}$. Accordingly, we have

$$|\Theta| \leq \prod_{i=1}^{l+1} n_\epsilon{}^i = \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1)+\sum_{i=1}^{l} a_i(k_i{}^2 d_i + d_{i-1} - 2a_i+1)} = \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{\theta},$$

where $\theta = a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1) + \sum_{i=1}^{l} a_i(k_i{}^2 d_i + d_{i-1} - 2a_i + 1)$ which is the total freedom degree of the network. So we can always find a vector $\boldsymbol{w}_{k_{\boldsymbol{w}}} \in \Theta$ such that $\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2 \leq \epsilon$. Now we use the decomposition strategy to bound our goal:

$$\begin{aligned}
&\left\|\nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla^2 \boldsymbol{Q}(\boldsymbol{w})\right\|_{\text{op}} \\
=&\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}))\right\|_{\text{op}} \\
=&\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)})\right) + \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}))\right. \\
&\left. + \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}))\right\|_{\text{op}} \\
\leq&\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)})\right)\right\|_{\text{op}} + \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}))\right\|_{\text{op}} \\
&+ \left\|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}))\right\|_{\text{op}}.
\end{aligned}$$

Here we also define four events $\boldsymbol{E}_0$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ and $\boldsymbol{E}_3$ as

$$\begin{aligned}
\boldsymbol{E}_0 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla^2 \boldsymbol{Q}(\boldsymbol{w})\right\|_{\text{op}} \geq t\right\}, \\
\boldsymbol{E}_1 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)})\right)\right\|_{\text{op}} \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_2 &= \left\{\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}))\right\|_{\text{op}} \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_3 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left\|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}))\right\|_{\text{op}} \geq \frac{t}{3}\right\}.
\end{aligned}$$

Accordingly, we have

$$\mathbb{P}\left(\boldsymbol{E}_0\right) \leq \mathbb{P}\left(\boldsymbol{E}_1\right) + \mathbb{P}\left(\boldsymbol{E}_2\right) + \mathbb{P}\left(\boldsymbol{E}_3\right).$$

So we can respectively bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$, $\mathbb{P}\left(\boldsymbol{E}_2\right)$ and $\mathbb{P}\left(\boldsymbol{E}_3\right)$ to bound $\mathbb{P}\left(\boldsymbol{E}_0\right)$.

Step 1. Bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_1\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w},\boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right\|_2 \geq \frac{t}{3}\right)\\
\overset{①}{\leq}& \frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w},\boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right\|_2\right)\\
\leq& \frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla^2 f(\boldsymbol{w},\boldsymbol{D}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D})\right\|_2\right)\\
\leq& \frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\frac{\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f(\boldsymbol{w},\boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right\|_2}{\left\|\boldsymbol{w}-\boldsymbol{w}_{k_{\boldsymbol{w}}}\right\|_2}\sup_{\boldsymbol{w}\in\Omega}\left\|\boldsymbol{w}-\boldsymbol{w}_{k_{\boldsymbol{w}}}\right\|_2\right)\\
\overset{②}{\leq}& \frac{3\nu\epsilon}{t},
\end{aligned}$$

where ① holds since by Markov inequality and ② holds because of Lemma 14. Therefore, we can set

$$t \geq \frac{6\nu\epsilon}{\varepsilon}.$$

Then we can bound $\mathbb{P}(\boldsymbol{E}_1)$:

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{\varepsilon}{2}.$$

Step 2. Bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$: By Lemma 3, we know that for any matrix $\boldsymbol{X}\in\mathbb{R}^{d\times d}$, its operator norm can be computed as

$$\|\boldsymbol{X}\|_{\text{op}} \leq \frac{1}{1-2\epsilon}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}_\epsilon}\left|\langle\boldsymbol{\lambda},\boldsymbol{X}\boldsymbol{\lambda}\rangle\right|.$$

where $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_{k_{\boldsymbol{w}}}\}$ be an ϵ-covering net of $\mathsf{B}^d(1)$.

Let $\boldsymbol{\lambda}_{1/4}$ be the $\frac{1}{4}$-covering net of $\mathsf{B}^d(1)$, where $d = \tilde{r}_l\tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1}d_i$. Recall that we use Θ to denote the ϵ-net of $\boldsymbol{w}_{k_{\boldsymbol{w}}}$ and we have $|\Theta| \leq \prod_{i=1}^{l+1} {n_\epsilon}^i = \left(\frac{3(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta$. Then we can bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_2\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))\right\|_2 \geq \frac{t}{3}\right)\\
\leq& \mathbb{P}\left(\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta,\boldsymbol{\lambda}\in\boldsymbol{\lambda}_{1/4}} 2\left|\left\langle\boldsymbol{\lambda},\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D})\right)\right)\boldsymbol{\lambda}\right\rangle\right| \geq \frac{t}{3}\right)\\
\leq& 12^d\left(\frac{3(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta \sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta,\boldsymbol{\lambda}\in\boldsymbol{\lambda}_{1/4}}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle\boldsymbol{\lambda},\left(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D})\right)\right)\boldsymbol{\lambda}\right\rangle\right| \geq \frac{t}{6}\right)\\
\overset{①}{\leq}& 12^d\left(\frac{3(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta 2\exp\left(-\frac{nt^2}{72\gamma^2}\right),
\end{aligned}$$

where ① holds since by Lemma 17, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle\boldsymbol{\lambda},(\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w},\boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w},\boldsymbol{D}^{(i)}))\boldsymbol{\lambda}\right\rangle\right) > t\right) \leq \exp\left(-\frac{nt^2}{2\gamma^2}\right).$$

where $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2} p^2} \right]^2 \right)^{1/2}$.

Thus, if we set

$$t \geq \sqrt{\frac{72\gamma^2 \left(d \log(12) + \theta \log\left(\frac{3(b_{l+1} + \sum_{i=1}^{l} d_i b_i)}{\epsilon} \right) + \log\left(\frac{4}{\varepsilon}\right) \right)}{n}},$$

then we have

$$\mathbb{P}\left(\boldsymbol{E}_2\right) \leq \frac{\varepsilon}{2}.$$

Step 3. Bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_3\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w} \in \Omega} \left\| \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} (\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} (\nabla^2 f(\boldsymbol{w}, \boldsymbol{D})) \right\|_2 \geq \frac{t}{3} \right) \\
\leq & \mathbb{P}\left(\mathbb{E}_{\boldsymbol{D} \sim \boldsymbol{\mathcal{D}}} \sup_{\boldsymbol{w} \in \Omega} \left\| (\nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}) - \nabla^2 f(\boldsymbol{w}, \boldsymbol{D}) \right\|_2 \geq \frac{t}{3} \right) \\
\leq & \mathbb{P}\left(\sup_{\boldsymbol{w} \in \Omega} \frac{\left| \frac{1}{n} \sum_{i=1}^{n} \left(\nabla^2 f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla^2 f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) \right) \right|}{\left\| \boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}} \right\|_2} \sup_{\boldsymbol{w} \in \Omega} \left\| \boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}} \right\|_2 \geq \frac{t}{3} \right) \\
\overset{①}{\leq} & \mathbb{P}\left(\nu \epsilon \geq \frac{t}{3} \right),
\end{aligned}$$

where ① holds because of Lemma 14. We set ϵ enough small such that $\nu\epsilon < t/3$ always holds. Then it yields $\mathbb{P}\left(\boldsymbol{E}_3\right) = 0$.

Step 4. Final result: To ensure $\mathbb{P}(\boldsymbol{E}_0) \leq \varepsilon$, we just set $\epsilon = \frac{36(b_{l+1} + \sum_{i=1}^{l} d_i b_i)}{\vartheta^2 n b_{l+1}} \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2} p^2} \right]^{-\frac{1}{2}}$. Note that $\frac{6\epsilon\nu}{\varepsilon} > 3\epsilon\nu$. Thus we can obtain

$$t \geq \max\left(\frac{6\nu\epsilon}{\varepsilon}, \sqrt{\frac{72\gamma^2 \left(d \log(12) + \theta \log\left(\frac{3(b_{l+1} + \sum_{i=1}^{l} d_i b_i)}{\epsilon} \right) + \log\left(\frac{4}{\varepsilon}\right) \right)}{n}} \right).$$

Thus, if $n \geq c_{v'} \frac{\nu^2}{d\varrho\varepsilon^2} \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2} p^2} \right]^{-1}$ where $c_{v'}$ is a constant, there exists a universal constant c_v such that

$$\begin{aligned}
\sup_{\boldsymbol{w} \in \Omega} \left\| \nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla^2 \boldsymbol{Q}(\boldsymbol{w}) \right\|_{\text{op}} \leq & \hat{c}_v \gamma \sqrt{\frac{d \log(12) + \theta \left(\sum_{i=1}^{l} \log\left(\frac{\sqrt{d_s} b_s (k_s - s_s + 1)}{4p} \right) + \log(b_{l+1}) + \log\left(\frac{n}{128 p^2} \right) \right) + \log\left(\frac{4}{\varepsilon}\right)}{n}} \\
= & c_v \gamma \sqrt{\frac{2d + \theta\varrho + \log\left(\frac{4}{\varepsilon}\right)}{2n}}
\end{aligned}$$

holds with probability at least $1 - \varepsilon$, where $d = \tilde{r}_l \tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, $\theta = a_{l+1}(d_{l+1} + \tilde{r}_l \tilde{c}_l d_l - 2a_{l+1} + 1) + \sum_{i=1}^{l} a_i({k_i}^2 d_i + d_{i-1} - 2a_i + 1)$, $\varrho = \sum_{i=1}^{l} \log\left(\frac{\sqrt{d_i} b_i (k_i - s_i + 1)}{4p} \right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2} \right)$, and $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2} p^2} \right]^2 \right)^{1/2}$. The proof is completed. □

D Proofs of Main Theorems

D.1 Proof of Lemma 1

Proof. Recall that the weight of each kernel and the feature maps has magnitude bound separately, *i.e.* $\boldsymbol{w}_{(i)}^k \in \mathsf{B}^{{k_i}^2 d_{i-1}}(r_w)$ $(i = 1, \cdots, l; k = 1, \cdots, d_i)$ and $\boldsymbol{w}_{(l+1)} \in \mathsf{B}^{\tilde{r}_l \tilde{c}_l d_l d_{l+1}}(b_{l+1})$. Since $\widetilde{\boldsymbol{W}}_{(i)} = [\mathsf{vec}(\boldsymbol{W}_{(i)}^1), \mathsf{vec}(\boldsymbol{W}_{(i)}^2), \cdots, \mathsf{vec}(\boldsymbol{W}_{(i)}^{d_{i-1}})] \in \mathbb{R}^{k_i^2 d_i \times d_{i-1}}$, we have $\|\widetilde{\boldsymbol{W}}_{(i)}\|_F \leq d_i b_i$.

So here we assume $\widetilde{\boldsymbol{W}}_{(i,\epsilon)}$ is the $d_i b_i \epsilon/(b_{l+1}+\sum_{i=1}^{l} d_i b_i)$-covering net of the matrix $\widetilde{\boldsymbol{W}}_{(i)}$ which is the set of all parameters in the i-th layer. Then by Lemma 7, we have the covering number

$$n_\epsilon{}^i \leq \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_i({k_i}^2 d_i + d_{i-1} - 2a_i + 1)},$$

since the rank of $\widetilde{\boldsymbol{W}}_{(i)}$ obeys $\mathsf{rank}(\widetilde{\boldsymbol{W}}_{(i)}) \leq a_i$ for $1 \leq i \leq l$. For the last layer, we also can construct an $b_{l+1}\epsilon/(b_{l+1}+\sum_{i=1}^{l} d_i b_i)$-covering net for the weight matrix $\boldsymbol{W}_{(l+1)}$. Here we have

$$n_\epsilon{}^{l+1} \leq \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1)},$$

since the rank of $\boldsymbol{W}_{(l+1)}$ obeys $\mathsf{rank}(\boldsymbol{W}_{(l+1)}) \leq a_{l+1}$. Finally, we arrange them together to construct a set Θ and claim that there is always an ϵ-covering net $\boldsymbol{w}_\epsilon$ in Θ for any parameter $\boldsymbol{w}$. Accordingly, we have

$$|\Theta| \leq \prod_{i=1}^{l+1} n_\epsilon{}^i = \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1)+\sum_{i=1}^{l} a_i({k_i}^2 d_i + d_{i-1} - 2a_i+1)} = \left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^{\theta},$$

where $\theta = a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_l d_l - 2a_{l+1}+1)+\sum_{i=1}^{l} a_i({k_i}^2 d_i + d_{i-1} - 2a_i + 1)$ which is the total freedom degree of the network. So we can always find a vector $\boldsymbol{w}_{k_{\boldsymbol{w}}} \in \Theta$ such that $\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2 \leq \epsilon$. Now we use the decomposition strategy to bound our goal:

$$\begin{aligned}
&\left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \boldsymbol{Q}(\boldsymbol{w})\right| = \left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}, \boldsymbol{D}))\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)})\Big) + \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}) + \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} f(\boldsymbol{w}, \boldsymbol{D})\right| \\
\leq&\left|\frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)})\Big)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})\right| \\
&+\left|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} f(\boldsymbol{w}, \boldsymbol{D})\right|.
\end{aligned}$$

Then, we define four events $\boldsymbol{E}_0$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ and $\boldsymbol{E}_3$ as

$$\begin{aligned}
\boldsymbol{E}_0 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \boldsymbol{Q}(\boldsymbol{w})\right| \geq t\right\}, \\
\boldsymbol{E}_1 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{x}_{(i)})\Big)\right| \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_2 &= \left\{\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}))\right| \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_3 &= \left\{\sup_{\boldsymbol{w}\in\Omega}\left|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}, \boldsymbol{D}))\right| \geq \frac{t}{3}\right\}.
\end{aligned}$$

Accordingly, we have

$$\mathbb{P}(\boldsymbol{E}_0) \leq \mathbb{P}(\boldsymbol{E}_1) + \mathbb{P}(\boldsymbol{E}_2) + \mathbb{P}(\boldsymbol{E}_3).$$

So we can respectively bound $\mathbb{P}(\boldsymbol{E}_1)$, $\mathbb{P}(\boldsymbol{E}_2)$ and $\mathbb{P}(\boldsymbol{E}_3)$ to bound $\mathbb{P}(\boldsymbol{E}_0)$.

Step 1. Bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$ as follows:

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{E}_1\right)=&\mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{w},\boldsymbol{D}^{(i)})-f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right|\geq\frac{t}{3}\right)\\ \overset{①}{\leq}&\frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{w},\boldsymbol{D}^{(i)})-f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right|\right)\\ \leq&\frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{w},\boldsymbol{D}^{(i)})-f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right|}{\left\|\boldsymbol{w}-\boldsymbol{w}_{k_{\boldsymbol{w}}}\right\|_2}\sup_{\boldsymbol{w}\in\Omega}\left\|\boldsymbol{w}-\boldsymbol{w}_{k_{\boldsymbol{w}}}\right\|_2\right)\\ \leq&\frac{3\epsilon}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w},\boldsymbol{D})\right\|_2\right),\end{aligned}$$

where ① holds since by Markov inequality, we have that for an arbitrary nonnegative random variable x, then

$$\mathbb{P}(x\geq t)\leq\frac{\mathbb{E}(x)}{t}.$$

Now we only need to bound $\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w},\boldsymbol{D})\right\|_2\right)$. Therefore, by Lemma 11, we have

$$\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w},\boldsymbol{D})\right\|_2\right)=\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(\boldsymbol{w},\boldsymbol{D}^{(i)})\right\|_2\right)\leq\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\|\nabla f(\boldsymbol{w},\boldsymbol{D})\|_2\right)\leq\beta.$$

where $\beta\triangleq\left[\vartheta\tilde{r}_l\tilde{c}_l d_l+\sum_{i=1}^{l}\frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2{b_i}^2 d_i}r_{i-1}c_{i-1}\prod_{s=i}^{l}\frac{d_s{b_s}^2(k_s-s_s+1)^2}{16p^2}\right]^{1/2}$ in which $\vartheta=1/8$. Therefore, we have

$$\mathbb{P}\left(\boldsymbol{E}_1\right)\leq\frac{3\epsilon\beta}{t}.$$

We further let

$$t\geq\frac{6\epsilon\beta}{\varepsilon}.$$

Then we can bound $\mathbb{P}(\boldsymbol{E}_1)$:

$$\mathbb{P}(\boldsymbol{E}_1)\leq\frac{\varepsilon}{2}.$$

Step 2. Bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$: Recall that we use Θ to denote the index of $\boldsymbol{w}_{k_{\boldsymbol{w}}}$ and we have $|\Theta|\leq\prod_{i=1}^{l+1}{n_\epsilon}^i=\left(\frac{9(b_{l+1}+\sum_{i=1}^{l}d_ib_i)}{\epsilon}\right)^\theta$. We can bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$ as follows:

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{E}_2\right)=&\mathbb{P}\left(\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))\right|\geq\frac{t}{3}\right)\\ \leq&\left(\frac{9(b_{l+1}+\sum_{i=1}^{l}d_ib_i)}{\epsilon}\right)^\theta\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))\right|\geq\frac{t}{3}\right)\\ \overset{①}{\leq}&\left(\frac{9(b_{l+1}+\sum_{i=1}^{l}d_ib_i)}{\epsilon}\right)^\theta 2\exp\left(-\frac{2nt^2}{\alpha^2}\right),\end{aligned}$$

where ① holds because in Lemma 15, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{w},\boldsymbol{D}^{(i)})-\mathbb{E}(f(\boldsymbol{w},\boldsymbol{D}^{(i)}))\right)>t\right)\leq\exp\left(-\frac{2nt^2}{\alpha^2}\right),$$

where $\alpha = 1$. Thus, if we set

$$t \geq \sqrt{\frac{\alpha^2 \left(\theta \log\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right) + \log\left(\frac{4}{\varepsilon}\right)\right)}{2n}},$$

then we have

$$\mathbb{P}\left(\boldsymbol{E}_2\right) \leq \frac{\varepsilon}{2}.$$

Step 3. Bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_3\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega} \|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(f(\boldsymbol{w}, \boldsymbol{D}))\|_2 \geq \frac{t}{3}\right) \\
=& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega} \frac{\|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}) - f(\boldsymbol{w}, \boldsymbol{D})\right)\|_2)}{\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2} \sup_{\boldsymbol{w}\in\Omega} \|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2 \geq \frac{t}{3}\right) \\
\leq& \mathbb{P}\left(\epsilon \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \sup_{\boldsymbol{w}\in\Omega} \|\nabla \boldsymbol{Q}_{\boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{D})\|_2 \geq \frac{t}{3}\right) \\
\overset{①}{\leq}& \mathbb{P}\left(\beta\epsilon \geq \frac{t}{3}\right),
\end{aligned}$$

where ① holds since we utilize Lemma 11. We set ϵ enough small such that $\beta\epsilon < t/3$ always holds. Then it yields $\mathbb{P}\left(\boldsymbol{E}_3\right) = 0$.

Step 4. Final result: To ensure $\mathbb{P}(\boldsymbol{E}_0) \leq \varepsilon$, we just set $\epsilon = \frac{18p^2(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\vartheta^2 n b_{l+1}} \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2}\right]^{-\frac{1}{2}}$. Note that $\frac{6\epsilon\beta}{\varepsilon} > 3\epsilon\beta$ due to $\varepsilon \leq 1$. Thus we can obtain

$$t \geq \max\left(\frac{6\epsilon\beta}{\varepsilon}, \sqrt{\frac{\alpha^2 \left(\theta \log\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right) + \log\left(\frac{4}{\varepsilon}\right)\right)}{2n}}\right).$$

By comparing the values of α, we can observe that if $n \geq c_{f'} \frac{l^2 (b_{l+1}+\sum_{i=1}^{l} d_i b_i)^2 \max_i \sqrt{r_i c_i}}{\theta \varrho \varepsilon^2}$ where $c_{f'}$ is a constant, there exists such a universal constant c_f such that

$$\sup_{\boldsymbol{w}\in\Omega} \left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \boldsymbol{Q}(\boldsymbol{w})\right| \leq \alpha \sqrt{\frac{\theta\left(\sum_{i=1}^{l} \log\left(\frac{\sqrt{d_s} b_s (k_s - s_s + 1)}{4p}\right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2}\right)\right) + \log\left(\frac{4}{\varepsilon}\right)}{2n}} = \sqrt{\frac{\theta\varrho + \log\left(\frac{4}{\varepsilon}\right)}{2n}}$$

holds with probability at least $1 - \varepsilon$, where $\theta = a_{l+1}(d_{l+1} + \tilde{r}_l \tilde{c}_l d_l - 2a_{l+1} + 1) + \sum_{i=1}^{l} a_i(k_i{}^2 d_i + d_{i-1} - 2a_i + 1)$, $\varrho = \sum_{i=1}^{l} \log\left(\frac{\sqrt{d_i} b_i (k_i - s_i + 1)}{4p}\right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2}\right)$, and $\alpha = 1$. The proof is completed. □

D.2 Proof of Theorem 1

Proof. By Lemma 1 in the manuscript, we know that if $n \geq c_{f'} l^2 (b_{l+1} + \sum_{i=1}^{l} d_i b_i)^2 \max_i \sqrt{r_i c_i}/(\theta\varrho\varepsilon^2)$ where $c_{f'}$ is a universal constant, then with probability at least $1 - \varepsilon$, we have

$$\sup_{\boldsymbol{w}\in\Omega} \left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \boldsymbol{Q}(\boldsymbol{w})\right| \leq \sqrt{\frac{\theta\varrho + \log\left(\frac{4}{\varepsilon}\right)}{2n}},$$

where the total freedom degree θ of the network is $\theta = a_{l+1}(d_{l+1} + \tilde{r}_l \tilde{c}_l d_l + 1) + \sum_{i=1}^{l} a_i(k_i{}^2 d_{i-1} + d_i + 1)$ and $\varrho = \sum_{i=1}^{l} \log\left(\frac{\sqrt{d_i} b_i (k_i - s_i + 1)}{4p}\right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2}\right)$.

Thus based on such a result, we can derive the following generalization bound:

$$\mathbb{E}_{\boldsymbol{\mathcal{S}}\sim\boldsymbol{\mathcal{D}}}\left|\mathbb{E}_{\mathcal{A}}(\boldsymbol{Q}(\widetilde{\boldsymbol{w}})-\widetilde{\boldsymbol{Q}}_n(\widetilde{\boldsymbol{w}}))\right|\leq\mathbb{E}_{\boldsymbol{\mathcal{S}}\sim\boldsymbol{\mathcal{D}}}\left(\sup_{\boldsymbol{w}\in\Omega}\left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\boldsymbol{Q}(\boldsymbol{w})\right|\right)\leq\sup_{\boldsymbol{w}\in\Omega}\left|\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\boldsymbol{Q}(\boldsymbol{w})\right|\leq\sqrt{\frac{\theta\varrho+\log\left(\frac{4}{\varepsilon}\right)}{2n}}.$$

Thus, the conclusion holds. The proof is completed. □

D.3 Proof of Theorem 2

Proof. Recall that the weight of each kernel and the feature maps has magnitude bound separately, *i.e.* $\boldsymbol{w}_{(i)}^k\in\mathsf{B}^{{k_i}^2d_{i-1}}(r_w)$ $(i=1,\cdots,l;k=1,\cdots,d_i)$ and $\boldsymbol{w}_{(l+1)}\in\mathsf{B}^{\tilde{r}_l\tilde{c}_ld_ld_{l+1}}(b_{l+1})$. Since $\widetilde{\boldsymbol{W}}_{(i)}=[\mathsf{vec}(\boldsymbol{W}_{(i)}^1),\mathsf{vec}(\boldsymbol{W}_{(i)}^2),\cdots,\mathsf{vec}(\boldsymbol{W}_{(i)}^{d_{i-1}})]\in\mathbb{R}^{k_i^2d_{i-1}\times d_i}$, we have $\|\widetilde{\boldsymbol{W}}_{(i)}\|_F\leq d_ib_i$.

So here we assume $\widetilde{\boldsymbol{W}}_{(i,\epsilon)}$ is the $d_ib_i\epsilon/(b_{l+1}+\sum_{i=1}^l d_ib_i)$-covering net of the matrix $\widetilde{\boldsymbol{W}}_{(i)}$ which is the set of all parameters in the i-th layer. Then by Lemma 7, we have the covering number

$${n_\epsilon}^i\leq\left(\frac{9(b_{l+1}+\sum_{i=1}^l d_ib_i)}{\epsilon}\right)^{a_i({k_i}^2d_{i-1}+d_i-2a_i+1)},$$

since the rank of $\widetilde{\boldsymbol{W}}_{(i)}$ obeys $\mathsf{rank}(\widetilde{\boldsymbol{W}}_{(i)})\leq a_i$ for $1\leq i\leq l$. For the last layer, we also can construct an $b_{l+1}\epsilon/(b_{l+1}+\sum_{i=1}^l d_ib_i)$-covering net for the weight matrix $\boldsymbol{W}_{(l+1)}$. Here we have

$${n_\epsilon}^{l+1}\leq\left(\frac{9(b_{l+1}+\sum_{i=1}^l d_ib_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_ld_l-2a_{l+1}+1)},$$

since the rank of $\boldsymbol{W}_{(l+1)}$ obeys $\mathsf{rank}(\boldsymbol{W}_{(l+1)})\leq a_{l+1}$. Finally, we arrange them together to construct a set Θ and claim that there is always an ϵ-covering net $\boldsymbol{w}_\epsilon$ in Θ for any parameter $\boldsymbol{w}$. Accordingly, we have

$$|\Theta|\leq\prod_{i=1}^{l+1}{n_\epsilon}^i=\left(\frac{9(b_{l+1}+\sum_{i=1}^l d_ib_i)}{\epsilon}\right)^{a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_ld_l-2a_{l+1}+1)+\sum_{i=1}^l a_i({k_i}^2d_{i-1}+d_i-2a_i+1)}=\left(\frac{9(b_{l+1}+\sum_{i=1}^l d_ib_i)}{\epsilon}\right)^{\theta},$$

where $\theta=a_{l+1}(d_{l+1}+\tilde{r}_l\tilde{c}_ld_l-2a_{l+1}+1)+\sum_{i=1}^l a_i({k_i}^2d_{i-1}+d_i-2a_i+1)$ which is the total freedom degree of the network. So we can always find a vector $\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta$ such that $\|\boldsymbol{w}-\boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2\leq\epsilon$. Accordingly, we can decompose $\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2$ as

$$\begin{aligned}
&\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2\\
=&\left\|\frac{1}{n}\sum_{i=1}^n\nabla f(\boldsymbol{w},\boldsymbol{D}^{(i)})-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w},\boldsymbol{D}))\right\|_2\\
=&\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f(\boldsymbol{w},\boldsymbol{D}^{(i)})-\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)+\frac{1}{n}\sum_{i=1}^n\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))\right.\\
&\left.+\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w},\boldsymbol{D}))\right\|_2\\
\leq&\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f(\boldsymbol{w},\boldsymbol{D}^{(i)})-\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})\right)\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^n\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}^{(i)})-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))\right\|_2\\
&+\left\|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}},\boldsymbol{D}))-\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w},\boldsymbol{D}))\right\|_2.
\end{aligned}$$

Here we also define four events $\boldsymbol{E}_0$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ and $\boldsymbol{E}_3$ as

$$\begin{aligned}
\boldsymbol{E}_0 &= \left\{ \sup_{\boldsymbol{w}\in\Omega} \left\| \nabla \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla \boldsymbol{Q}(\boldsymbol{w}) \right\|_2 \geq t \right\}, \\
\boldsymbol{E}_1 &= \left\{ \sup_{\boldsymbol{w}\in\Omega} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) \right) \right\|_2 \geq \frac{t}{3} \right\}, \\
\boldsymbol{E}_2 &= \left\{ \sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) \right\|_2 \geq \frac{t}{3} \right\}, \\
\boldsymbol{E}_3 &= \left\{ \sup_{\boldsymbol{w}\in\Omega} \left\| \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}, \boldsymbol{D})) \right\|_2 \geq \frac{t}{3} \right\}.
\end{aligned}$$

Accordingly, we have

$$\mathbb{P}\left(\boldsymbol{E}_0\right) \leq \mathbb{P}\left(\boldsymbol{E}_1\right) + \mathbb{P}\left(\boldsymbol{E}_2\right) + \mathbb{P}\left(\boldsymbol{E}_3\right).$$

So we can respectively bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$, $\mathbb{P}\left(\boldsymbol{E}_2\right)$ and $\mathbb{P}\left(\boldsymbol{E}_3\right)$ to bound $\mathbb{P}\left(\boldsymbol{E}_0\right)$.

Step 1. Bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_1\right)$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_1\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) \right) \right\|_2 \geq \frac{t}{3} \right) \\
\overset{①}{\leq}& \frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) \right) \right\|_2 \right) \\
\leq& \frac{3}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \frac{\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) \right) \right\|_2}{\left\| \boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}} \right\|_2} \sup_{\boldsymbol{w}\in\Omega} \left\| \boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}} \right\|_2 \right) \\
\leq& \frac{3\epsilon}{t}\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \left\| \nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}, \boldsymbol{D}) \right\|_2 \right),
\end{aligned}$$

where ① holds since by Markov inequality, we have that for an arbitrary nonnegative random variable x, then $\mathbb{P}(x \geq t) \leq \frac{\mathbb{E}(x)}{t}$.

Now we only need to bound $\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \left\| \nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}, \boldsymbol{D}) \right\|_2 \right)$. Here we utilize Lemma 14 to achieve this goal:

$$\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \left\| \nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}, \boldsymbol{D}) \right\|_2 \right) \leq \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}} \left(\sup_{\boldsymbol{w}\in\Omega} \left\| \nabla^2 f(\boldsymbol{w}, \boldsymbol{D}) - \nabla^2 f(\boldsymbol{w}^*, \boldsymbol{D}) \right\|_2 \right) \leq \gamma.$$

where $\gamma = \left(\frac{\vartheta {b_{l+1}}^2 d_0^2}{{b_1}^4 d_1^2} l^2 r_0^2 c_0^2 \left[\prod_{s=1}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{8\sqrt{2}p^2} \right]^2 \right)^{1/2}$. Therefore, we have

$$\mathbb{P}\left(\boldsymbol{E}_1\right) \leq \frac{3\gamma\epsilon}{t}.$$

We further let

$$t \geq \frac{6\gamma\epsilon}{\varepsilon}.$$

Then we can bound $\mathbb{P}(\boldsymbol{E}_1)$:

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{\varepsilon}{2}.$$

Step 2. Bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$: By Lemma 2, we know that for any vector $\boldsymbol{x} \in \mathbb{R}^d$, its ℓ_2-norm can be computed as

$$\|\boldsymbol{x}\|_2 \leq \frac{1}{1-\epsilon} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}_\epsilon} \langle \boldsymbol{\lambda}, \boldsymbol{x} \rangle.$$

where $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k_{\boldsymbol{w}}}\}$ be an ϵ-covering net of $\mathsf{B}^d(1)$.

Let $\boldsymbol{\lambda}$ be the $\frac{1}{2}$-covering net of $\mathsf{B}^d(1)$, where $d = \tilde{r}_l\tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$. Recall that we use Θ to denote the index of $\boldsymbol{w}_{k_{\boldsymbol{w}}}$ so that $\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\| \leq \epsilon$. Besides, $|\Theta| \leq \prod_{i=1}^{l+1} {n_\epsilon}^i = \left(\frac{3(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta$. Then we can bound $\mathbb{P}\left(\boldsymbol{E}_2\right)$ as follows:

$$
\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_2\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}))\right\|_2 \geq \frac{t}{3}\right) \\
=& \mathbb{P}\left(\sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta, \boldsymbol{\lambda}\in\boldsymbol{\lambda}_{1/2}} 2\left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n}\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})\right)\right\rangle \geq \frac{t}{3}\right) \\
\leq& 6^d\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta \sup_{\boldsymbol{w}_{k_{\boldsymbol{w}}}\in\Theta, \boldsymbol{\lambda}\in\boldsymbol{\lambda}_{1/2}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left\langle \boldsymbol{\lambda}, \nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{D})\right)\right\rangle \geq \frac{t}{6}\right) \\
\overset{①}{\leq}& 6^d\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right)^\theta 2\exp\left(-\frac{nt^2}{72\beta^2}\right),
\end{aligned}
$$

where ① holds since by Lemma 16, we have

$$
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)}) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{D}^{(i)})\right\rangle\right) > t\right) \leq \exp\left(-\frac{nt^2}{2\beta^2}\right).
$$

where $\beta \triangleq \left[\vartheta\tilde{r}_l\tilde{c}_l d_l + \sum_{i=1}^{l} \frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1} c_{i-1} \prod_{s=i}^{l} \frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2}\right]^{1/2}$ in which $\vartheta = 1/8$.

Thus, if we set

$$
t \geq \sqrt{\frac{72\beta^2\left(d\log(6) + \theta\log\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right) + \log\left(\frac{4}{\varepsilon}\right)\right)}{n}},
$$

then we have

$$
\mathbb{P}\left(\boldsymbol{E}_2\right) \leq \frac{\varepsilon}{2}.
$$

Step 3. Bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$: We first bound $\mathbb{P}\left(\boldsymbol{E}_3\right)$ as follows:

$$
\begin{aligned}
\mathbb{P}\left(\boldsymbol{E}_3\right) =& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega}\|\mathbb{E}(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{x})) - \mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}(\nabla f(\boldsymbol{w}, \boldsymbol{x}))\|_2 \geq \frac{t}{3}\right) \\
=& \mathbb{P}\left(\sup_{\boldsymbol{w}\in\Omega}\frac{\|\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\left(\nabla f(\boldsymbol{w}_{k_{\boldsymbol{w}}}, \boldsymbol{x}) - \nabla f(\boldsymbol{w}, \boldsymbol{x})\right)\|_2)}{\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2}\sup_{\boldsymbol{w}\in\Omega}\|\boldsymbol{w} - \boldsymbol{w}_{k_{\boldsymbol{w}}}\|_2 \geq \frac{t}{3}\right) \\
\leq& \mathbb{P}\left(\epsilon\mathbb{E}_{\boldsymbol{D}\sim\boldsymbol{\mathcal{D}}}\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}, \boldsymbol{x})\right\|_2 \geq \frac{t}{3}\right) \\
\leq& \mathbb{P}\left(\gamma\epsilon \geq \frac{t}{3}\right).
\end{aligned}
$$

We set ϵ enough small such that $\gamma\epsilon < t/3$ always holds. Then it yields $\mathbb{P}\left(\boldsymbol{E}_3\right) = 0$.

Step 4. Final result: Note that $\frac{6\beta\epsilon}{\varepsilon} \geq 3\beta\epsilon$. Finally, to ensure $\mathbb{P}(\boldsymbol{E}_0) \leq \varepsilon$, we just set $\epsilon = \frac{18p^2(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\vartheta^2 n b_{l+1}}\left[\prod_{s=i}^{l}\frac{d_s {b_s}^2 (k_s - s_s + 1)^2}{16p^2}\right]^{-\frac{1}{2}}$.

$$
t \geq \max\left(\frac{6\gamma\epsilon}{\varepsilon}, \sqrt{\frac{72\beta^2\left(d\log(6) + \theta\log\left(\frac{9(b_{l+1}+\sum_{i=1}^{l} d_i b_i)}{\epsilon}\right) + \log\left(\frac{4}{\varepsilon}\right)\right)}{n}}\right).
$$

By comparing the values of β and γ, we have if $n \geq c_{g'} \frac{l^2 {b_{l+1}}^2 (b_{l+1}+\sum_{i=1}^{l} d_i b_i)^2 (r_0 c_0 d_0)^4}{d_0^4 {b_1}^8 (d\log(6)+\theta\varrho)\varepsilon^2 \max_i(r_i c_i)}$ where $c_{g'}$ is a universal constant, then there exists a universal constant c_g such that

$$\sup_{\boldsymbol{w}\in\Omega} \left\|\nabla_{\boldsymbol{w}}\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla_{\boldsymbol{w}}\boldsymbol{Q}(\boldsymbol{w})\right\|_2 \leq c_g\beta\sqrt{\frac{d+\frac{1}{\log(6)}\theta\left[\left(\sum_{i=1}^{l}\log\left(\frac{\sqrt{d_s}b_s(k_s-s_s+1)}{4p}\right)+\log(b_{l+1})+\log\left(\frac{n}{128p^2}\right)\right)+\log\left(\frac{4}{\varepsilon}\right)\right]}{n}}$$
$$\leq c_g\beta\sqrt{\frac{d+\frac{1}{2}\theta\varrho+\frac{1}{2}\log\left(\frac{4}{\varepsilon}\right)}{n}},$$

holds with probability at least $1-\varepsilon$, where $d = \tilde{r}_l\tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, $\theta = a_{l+1}(d_{l+1} + \tilde{r}_l\tilde{c}_l d_l + 1) + \sum_{i=1}^{l} a_i({k_i}^2 d_{i-1} + d_i + 1)$, $\varrho = \sum_{i=1}^{l}\log\left(\frac{\sqrt{d_i}b_i(k_i-s_i+1)}{4p}\right) + \log(b_{l+1}) + \log\left(\frac{n}{128p^2}\right)$, and $\beta \triangleq \left[\vartheta\tilde{r}_l\tilde{c}_l d_l + \sum_{i=1}^{l}\frac{\vartheta {b_{l+1}}^2 d_{i-1}}{p^2 {b_i}^2 d_i} r_{i-1}c_{i-1}\prod_{s=i}^{l}\frac{d_s {b_s}^2 (k_s-s_s+1)^2}{16p^2}\right]^{1/2}$ in which $\vartheta = 1/8$. The proof is completed. □

D.4 Proof of Corollary 1

Proof. By Theorem 2, we know that there exist universal constants $c_{g'}$ and c_g such that if $n \geq c_{g'} \frac{l^2 {b_{l+1}}^2 (b_{l+1}+\sum_{i=1}^{l} d_i b_i)^2 (r_0 c_0 d_0)^4}{d_0^4 {b_1}^8 (d\log(6)+\theta\varrho)\varepsilon^2 \max_i(r_i c_i)}$, then

$$\sup_{\boldsymbol{w}\in\Omega} \left\|\nabla_{\boldsymbol{w}}\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) - \nabla_{\boldsymbol{w}}\boldsymbol{Q}(\boldsymbol{w})\right\|_2 \leq c_g\beta\sqrt{\frac{2d+\theta\varrho+\log\left(\frac{4}{\varepsilon}\right)}{2n}}$$

holds with probability at least $1-\varepsilon$, where ϱ is provided in Lemma 1. Here β and d are defined as $\beta = \left[\frac{r_l c_l d_l}{8p^2} + \sum_{i=1}^{l}\frac{{b_{l+1}}^2 d_{i-1}}{8p^2 {b_i}^2 d_i} r_{i-1}c_{i-1}\prod_{j=i}^{l}\frac{d_j {b_j}^2 (k_j-s_j+1)^2}{16p^2}\right]^{1/2}$ and $d = \tilde{r}_l\tilde{c}_l d_l d_{l+1} + \sum_{i=1}^{l} {k_i}^2 d_{i-1} d_i$, respectively.

So based on such a result, we can derive that if $n \geq c_g^2(2d+\theta\varrho+\log(4/\varepsilon))\beta^2/(2\epsilon)$, then we have

$$\|\nabla\boldsymbol{Q}(\widetilde{\boldsymbol{w}})\|_2 \leq \left\|\nabla_{\boldsymbol{w}}\widetilde{\boldsymbol{Q}}_n(\widetilde{\boldsymbol{w}})\right\|_2 + \left\|\nabla_{\boldsymbol{w}}\widetilde{\boldsymbol{Q}}_n(\widetilde{\boldsymbol{w}}) - \nabla_{\boldsymbol{w}}\boldsymbol{Q}(\widetilde{\boldsymbol{w}})\right\|_2 \leq \sqrt{\epsilon} + c_g\beta\sqrt{\frac{2d+\theta\varrho+\log\left(\frac{4}{\varepsilon}\right)}{2n}} \leq 2\sqrt{\epsilon}.$$

Thus, we have $\|\nabla\boldsymbol{Q}(\widetilde{\boldsymbol{w}})\|_2^2 \leq 4\epsilon$, which means that $\widetilde{\boldsymbol{w}}$ is a 4ϵ-approximate stationary point in population risk with probability at least $1-\varepsilon$. The proof is completed. □

D.5 Proof of Theorem 3

Proof. Suppose that $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \cdots, \boldsymbol{w}_{(m)}\}$ are the non-degenerate critical points of $\boldsymbol{Q}(\boldsymbol{w})$. So for any $\boldsymbol{w}_{(k)}$, it obeys

$$\inf_i \left|\lambda_i^k\left(\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)\right| \geq \zeta,$$

where $\lambda_i^k\left(\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)$ denotes the i-th eigenvalue of the Hessian $\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)})$ and ζ is a constant. We further define a set $D = \{\boldsymbol{w}\in\mathbb{R}^d \,|\, \|\nabla\boldsymbol{Q}(\boldsymbol{w})\|_2 \leq \epsilon \text{ and } \inf_i|\lambda_i\left(\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)| \geq \zeta\}$. According to Lemma 5, $D = \cup_{k=1}^{\infty} D_k$ where each D_k is a disjoint component with $\boldsymbol{w}_{(k)} \in D_k$ for $k \leq m$ and D_k does not contain any critical point of $\boldsymbol{Q}(\boldsymbol{w})$ for $k \geq m+1$. On the other hand, by the continuity of $\nabla\boldsymbol{Q}(\boldsymbol{w})$, it yields $\|\nabla\boldsymbol{Q}(\boldsymbol{w})\|_2 = \epsilon$ for $\boldsymbol{w} \in \partial D_k$. Notice, we set the value of ϵ blow which is actually a function related to n.

Then by utilizing Theorem 2, we let sample number n sufficient large such that

$$\sup_{\boldsymbol{w}\in\Omega} \left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2 \leq \frac{\epsilon}{2}$$

holds with probability at least $1-\varepsilon$, where ϵ is defined as

$$\frac{\epsilon}{2} \triangleq c_g\beta\sqrt{\frac{d\log(6)+\theta\left(\sum_{i=1}^{l}\log\left(\frac{\sqrt{d_s}b_s(k_s-s_s+1)}{4p}\right)+\log(b_{l+1})+\log\left(\frac{n}{128p^2}\right)\right)+\log\left(\frac{4}{\varepsilon}\right)}{n}}.$$

This further gives that for arbitrary $\boldsymbol{w} \in D_k$, we have

$$\begin{aligned}\inf_{\boldsymbol{w}\in D_k}\left\|t\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})+(1-t)\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2 &= \inf_{\boldsymbol{w}\in D_k}\left\|t\left(\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla\boldsymbol{Q}(\boldsymbol{w})\right)+\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2\\ &\geq \inf_{\boldsymbol{w}\in D_k}\|\nabla\boldsymbol{Q}(\boldsymbol{w})\|_2-\sup_{\boldsymbol{w}\in D_k} t\left\|\nabla\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla\boldsymbol{Q}(\boldsymbol{w})\right\|_2\\ &\geq \frac{\epsilon}{2}.\end{aligned} \tag{6}$$

Similarly, by utilizing Lemma 18, let n be sufficient large such that

$$\sup_{\boldsymbol{w}\in\Omega}\left\|\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right\|_{\text{op}}\leq\frac{\zeta}{2}$$

holds with probability at least $1-\varepsilon$, where ζ satisfies

$$\frac{\zeta}{2}\geq c_v\gamma\sqrt{\frac{d+\theta\varrho+\log\left(\frac{4}{\varepsilon}\right)}{n}}.$$

Assume that $\boldsymbol{b}\in\mathbb{R}^d$ is a vector and satisfies $\boldsymbol{b}^T\boldsymbol{b}=1$. In this case, we can bound $\lambda_i^k\left(\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})\right)$ for arbitrary $\boldsymbol{w}\in D_k$ as follows:

$$\begin{aligned}\inf_{\boldsymbol{w}\in D_k}\left|\lambda_i^k\left(\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})\right)\right| &= \inf_{\boldsymbol{w}\in D_k}\min_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})\boldsymbol{b}\right|\\ &= \inf_{\boldsymbol{w}\in D_k}\min_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\left(\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right)\boldsymbol{b}+\boldsymbol{b}^T\nabla^2\boldsymbol{Q}(\boldsymbol{w})\boldsymbol{b}\right|\\ &\geq \inf_{\boldsymbol{w}\in D_k}\min_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\nabla^2\boldsymbol{Q}(\boldsymbol{w})\boldsymbol{b}\right|-\min_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\left(\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right)\boldsymbol{b}\right|\\ &\geq \inf_{\boldsymbol{w}\in D_k}\min_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\nabla^2\boldsymbol{Q}(\boldsymbol{w})\boldsymbol{b}\right|-\max_{\boldsymbol{b}^T\boldsymbol{b}=1}\left|\boldsymbol{b}^T\left(\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right)\boldsymbol{b}\right|\\ &= \inf_{\boldsymbol{w}\in D_k}\inf_i|\lambda_i^k\left(\nabla^2 f(\boldsymbol{w}_{(k)},\boldsymbol{x})\right)|-\left\|\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})-\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right\|_{\text{op}}\\ &\geq \frac{\zeta}{2}.\end{aligned} \tag{7}$$

This means that in each set D_k, $\nabla^2\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$ has no zero eigenvalues. Then, combine this and Eqn. (6), by Lemma 4 we know that if the population risk $\boldsymbol{Q}(\boldsymbol{w})$ has no critical point in D_k, then the empirical risk $\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$ has also no critical point in D_k; otherwise it also holds.

Now we bound the distance between the corresponding critical points of $\boldsymbol{Q}(\boldsymbol{w})$ and $\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$. Assume that in D_k, $\boldsymbol{Q}(\boldsymbol{w})$ has a unique critical point $\boldsymbol{w}_{(k)}$ and $\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$ also has a unique critical point $\boldsymbol{w}_n^{(k)}$. Then, there exists $t\in[0,1]$ such that for any $\boldsymbol{z}\in\partial\mathsf{B}^d(1)$, we have

$$\begin{aligned}\epsilon \geq&\|\nabla\boldsymbol{Q}(\boldsymbol{w}_n^{(k)})\|_2\\ =&\max_{\boldsymbol{z}^T\boldsymbol{z}=1}\langle\nabla\boldsymbol{Q}(\boldsymbol{w}_n^{(k)}),\boldsymbol{z}\rangle\\ =&\max_{\boldsymbol{z}^T\boldsymbol{z}=1}\langle\nabla\boldsymbol{Q}(\boldsymbol{w}_{(k)}),\boldsymbol{z}\rangle+\langle\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)}+t(\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)}))(\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)}),\boldsymbol{z}\rangle\\ \overset{①}{\geq}&\left\langle\left(\nabla^2\boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)^2(\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)}),(\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)})\right\rangle^{1/2}\\ \overset{②}{\geq}&\zeta\|\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)}\|_2,\end{aligned}$$

where ① holds since $\nabla\boldsymbol{Q}(\boldsymbol{w}_{(k)})=\boldsymbol{0}$ and ② holds since $\boldsymbol{w}_{(k)}+t(\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)})$ is in D_k and for any $\boldsymbol{w}\in D_k$ we have $\inf_i|\lambda_i\left(\nabla^2\boldsymbol{Q}(\boldsymbol{w})\right)|\geq\zeta$. So if $n\geq c_h\max\left(\frac{l^2{b_{l+1}}^2(b_{l+1}+\sum_{i=1}^l d_ib_i)^2(r_0c_0d_0)^4}{d_0^4{b_1}^8d\varrho\varepsilon^2\max_i(r_ic_i)},\frac{d+\theta\varrho}{\zeta^2}\right)$ where c_h is a constant, then

$$\|\boldsymbol{w}_n^{(k)}-\boldsymbol{w}_{(k)}\|_2\leq\frac{2c_g\beta}{\zeta}\sqrt{\frac{d\log(6)+\theta\left(\sum_{i=1}^l\log\left(\frac{\sqrt{d_s}b_s(k_s-s_s+1)}{4p}\right)+\log(b_{l+1})+\log\left(\frac{n}{128p^2}\right)\right)+\log\left(\frac{4}{\varepsilon}\right)}{n}}$$

holds with probability at least $1-\varepsilon$. □

D.6 Proof of Corollary 2

Proof. By Theorem 3, we know that the non-degenerate stationary point $\boldsymbol{w}_{(k)}$ in the m non-degenerate stationary points in population risk, denoted by $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \cdots, \boldsymbol{w}_{(m)}\}$ uniquely corresponding to a non-degenerate stationary point $\boldsymbol{w}_n^{(k)}$ in the empirical risk.

On the other hand, for any $\boldsymbol{w}_{(k)}$, it obeys

$$\inf_i \left|\lambda_i^k\left(\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)\right| \geq \zeta,$$

where $\lambda_i^k\left(\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)$ denotes the i-th eigenvalue of the Hessian $\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_{(k)})$ and ζ is a constant. We further define a set $D = \{\boldsymbol{w} \in \mathbb{R}^d \,|\, \|\nabla \boldsymbol{Q}(\boldsymbol{w})\|_2 \leq \epsilon \text{ and } \inf_i |\lambda_i\left(\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_{(k)})\right)| \geq \zeta\}$. According to Lemma 5, $D = \cup_{k=1}^{\infty} D_k$ where each D_k is a disjoint component with $\boldsymbol{w}_{(k)} \in D_k$ for $k \leq m$ and D_k does not contain any critical point of $\boldsymbol{Q}(\boldsymbol{w})$ for $k \geq m+1$. Then $\boldsymbol{w}_n^{(k)}$ also belong to the component D_k due to the unique corresponding relation between $\boldsymbol{w}_{(k)}$ and $\boldsymbol{w}_n^{(k)}$. Then from Eqn. (6) and (7), we know that if the assumptions in Theorem 3 hold, then for arbitrary $\boldsymbol{w} \in D_k$ and $t \in (0,1)$,

$$\inf_{\boldsymbol{w}\in D_k} \left\|t\nabla \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w}) + (1-t)\nabla \boldsymbol{Q}(\boldsymbol{w})\right\|_2 \geq \frac{\epsilon}{2} \quad \text{and} \quad \inf_{\boldsymbol{w}\in D_k} \left|\lambda_i^k\left(\nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})\right)\right| \geq \frac{\zeta}{2},$$

where ϵ and ζ are constants. This means that in each set D_k, $\nabla^2 \widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$ has no zero eigenvalues. Then, combine this and Eqn. (6), we can obtain that in D_k, if $\boldsymbol{Q}(\boldsymbol{w})$ has a unique critical point $\boldsymbol{w}_{(k)}$ with non-degenerate index s_k, then $\widetilde{\boldsymbol{Q}}_n(\boldsymbol{w})$ also has a unique critical point $\boldsymbol{w}_{(k)}^n$ in D_k with the same non-degenerate index s_k. Namely, the number of negative eigenvalues of the Hessian matrices $\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_{(k)})$ and $\nabla^2 \boldsymbol{Q}(\boldsymbol{w}_n^{(k)})$ are the same. This further gives that if one of the pair $(\boldsymbol{w}_{(k)}, \boldsymbol{w}_n^{(k)})$ is a local minimum or saddle point, then another one is also a local minimum or a saddle point. The proof is completed. □

E Proof of Auxiliary Lemmas

E.1 Proof of Lemma 8

Proof. (1) Since $\mathsf{G}(\boldsymbol{z})$ is a diagonal matrix and its diagonal values are upper bounded by $\sigma_1(\boldsymbol{z}_i)(1-\sigma_1(\boldsymbol{z})) \leq 1/4$ where $\boldsymbol{z}_i$ denotes the i-th entry of $\boldsymbol{z}_i$, we can conclude

$$\|\mathsf{G}(\boldsymbol{z})\boldsymbol{M}\|_F^2 \leq \frac{1}{16}\|\boldsymbol{M}\|_F^2 \quad \text{and} \quad \|\boldsymbol{N}\mathsf{G}(\boldsymbol{z})\|_F^2 \leq \frac{1}{16}\|\boldsymbol{N}\|_F^2.$$

(2) The operator $\mathsf{Q}(\cdot)$ maps a vector $\boldsymbol{z} \in \mathbb{R}^d$ into a matrix of size $d^2 \times d$ whose $((i-1)d+i, i)$ $(i = 1, \cdots, d)$ entry equal to $\sigma_1(\boldsymbol{z}_i)(1-\sigma_1(\boldsymbol{z}_i))(1-2\sigma_1(\boldsymbol{z}_i))$ and rest entries are all 0. This gives

$$\begin{aligned}\sigma_1(\boldsymbol{z}_i)(1-\sigma_1(\boldsymbol{z}_i))(1-2\sigma_1(\boldsymbol{z}_i)) =& \frac{1}{3}(3\sigma_1(\boldsymbol{z}_i))(1-\sigma_1(\boldsymbol{z}_i))(1-2\sigma_1(\boldsymbol{z}_i)) \\ \leq& \frac{1}{3}\left(\frac{3\sigma_1(\boldsymbol{z}_i)+1-\sigma_1(\boldsymbol{z}_i)+1-2\sigma_1(\boldsymbol{z}_i)}{3}\right)^3 \\ \leq& \frac{2^3}{3^4}.\end{aligned}$$

This means the maximal value in $\mathsf{Q}(\boldsymbol{z})$ is at most $\frac{2^3}{3^4}$. Consider the structure in $\mathsf{Q}(\boldsymbol{z})$, we can obtain

$$\|\mathsf{Q}(\boldsymbol{z})\boldsymbol{M}\|_F^2 \leq \frac{2^6}{3^8}\|\boldsymbol{M}\|_F^2 \quad \text{and} \quad \|\boldsymbol{N}\mathsf{Q}(\boldsymbol{z})\|_F^2 \leq \frac{2^6}{3^8}\|\boldsymbol{N}\|_F^2.$$

(3) $\mathsf{up}(\boldsymbol{M})$ represents conducting upsampling on $\boldsymbol{M} \in \mathbb{R}^{s\times t\times q}$. Let $\boldsymbol{N} = \mathsf{up}(\boldsymbol{M}) \in \mathbb{R}^{ps\times pt\times q}$. Specifically, for each slice $\boldsymbol{N}(:,:,i)$ $(i = 1, \cdots, q)$, we have $\boldsymbol{N}(:,:,i) = \mathsf{up}(\boldsymbol{M}(:,:,i))$. It actually upsamples each entry $\boldsymbol{M}(g,h,i)$ into a matrix of p^2 same entries $\frac{1}{p^2}\boldsymbol{M}(g,h,i)$. So it is easy to obtain

$$\|\mathsf{up}(\boldsymbol{M})\|_F^2 \leq \frac{1}{p^2}\|\boldsymbol{M}\|_F^2.$$

(4) Let $\boldsymbol{M} = \boldsymbol{W}(:,:,i)$ and $\boldsymbol{N} = \widetilde{\boldsymbol{\delta}}_{i+1}(:,:i)$. Assume that $\boldsymbol{H} = \boldsymbol{M} \circledast \boldsymbol{N} \in \mathbb{R}^{m_1 \times m_2}$, where $m_1 = \tilde{r}_{i-1} - 2k_i + 2$ and $m_2 = \tilde{c}_{i-1} - 2k_i + 2$. Then we have

$$\|\boldsymbol{H}\|_F^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} |\boldsymbol{H}(i,j)|^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \langle \boldsymbol{M}_{\Omega_{i,j}}, \boldsymbol{N}\rangle^2 \leq \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \|\boldsymbol{M}_{\Omega_{i,j}}\|_F^2 \|\boldsymbol{N}\|_F^2,$$

where $\Omega_{i,j}$ denotes the entry index of $\boldsymbol{M}$ for the (i,j)-th convolution operation (*i.e.* computing the $\boldsymbol{H}(i,j)$).

Since for each convolution computing, each element in $\boldsymbol{M}$ is involved at most one time, we can claim that any element in $\boldsymbol{M}$ in $\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \|\boldsymbol{M}_{\Omega_{i,j}}\|_F^2$ occurs at most $(k_i - s_i + 1)^2$ since there are $s_i - 1$ rows and columns between each neighboring nonzero entries in $\boldsymbol{N}$ which is decided by the definition of $\widetilde{\boldsymbol{\delta}}_{i+1}$ in Sec. B.1. Therefore, we have

$$\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \|\boldsymbol{M}_{\Omega_{i,j}}\|_F^2 \leq (k_i - s_i + 1)^2 \|\boldsymbol{M}\|_F^2,$$

which further gives

$$\|\boldsymbol{M} \widetilde{\circledast} \boldsymbol{N}\|_F^2 \leq (k_i - s_i + 1)^2 \|\boldsymbol{M}\|_F^2 \|\boldsymbol{N}\|_F^2.$$

Consider all the slices in $\widetilde{\boldsymbol{\delta}}_{i+1}$, we can obtain

$$\|\widetilde{\boldsymbol{\delta}}_{i+1} \widetilde{\circledast} \boldsymbol{W}\|_F^2 \leq (k_i - s_i + 1)^2 \|\boldsymbol{W}\|_F^2 \|\widetilde{\boldsymbol{\delta}}_{i+1}\|_F^2.$$

(5) Since for softmax activation function σ_2, we have $\sum_{i=1}^{d_{l+1}} \boldsymbol{v}_i = 1$ ($\boldsymbol{v}_i \geq 0$) and there is only one nonzero entry (*i.e.* 1) in $\boldsymbol{y}$, we can obtain

$$0 \leq \|\boldsymbol{v} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{v}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{v}, \boldsymbol{y}\rangle = 2 - 2\langle \boldsymbol{v}, \boldsymbol{y}\rangle \leq 2.$$

The proof is completed. □

References

Candes, E. and Plan, Y. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE TIT*, 57(4):2342–2359, 2009.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301): 13–30, 1963.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for non-convex losses. *Annals of Statistics*, 2017.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices, compressed sensing. *Cambridge Univ. Press, Cambridge*, pp. 210–268, 2012.