ADMM-Enhanced CNN Training Framework With Global Convergence Guarantees

Chenjie Song, Zhengmin Kong, Shuo Liu, Li Ding, Tao Huang and Wei Xiang

I. Appendix

A. General Convergence of ADMM Without Normalization

Assumption 1: In this appendix, we consider more general settings than before, while X and Y are not necessarily normalized with unit norms. Before presenting our main theorem under these generic settings, we define the following constants:

$$\theta = \frac{LA}{L},\ \phi = (L_1\gamma)^{2(i-LA)}, \tag{1}$$

$$\alpha_1 = (L_1^2 + 2L_3C_3\gamma^i)\gamma^2, \tag{2}$$

$$\alpha_2 = (L_1^2 + 2L_3C_3\gamma^i + L_2C_3\gamma^i)\gamma^2, \tag{3}$$

$$\tau = 16(1+\phi) + 1, \tag{4}$$

$$L_3 = 2(L_1^2 + L_2L_0 + L_2), \tag{5}$$

$$\gamma = \max_{1\geq i\geq L} \|W_i^0\|_F, \tag{6}$$

$$m_{\min} = \min_{1\geq i\geq L-1} m_i, \tag{7}$$

$$f_{\min} = \sqrt{6}(\sqrt{3L_1} + 2(L_0L_3)^{1/2}(nm_{\min})^{1/4}), \tag{8}$$

$$\alpha_3 = (\frac{f_{\min}}{L_1})^2, \tag{9}$$

$$C_3 = \max\{\max_{0\geq i\geq L-2}\frac{2L_0\sqrt{nm_{j+1}}}{\gamma^j}, \frac{\|Y\|_F}{(\beta_L-3)\gamma^{L-1}}\}, \tag{10}$$

$$\lambda_1 = 3L_1C_3\beta_i\gamma^{i-3}(4C_3\gamma^{i-1} + L_0\sqrt{nm_l})(1 + \sqrt{\frac{6L_3C_3^2\gamma^{2(i-2)}}{L_1(4C_3\gamma^{i-1} + L_0\sqrt{nm_i})}}),\ 2\leq i\leq L-1 \tag{11}$$

$$\lambda_2 = \frac{1}{6}(1+3L_1^{-1}L_3\gamma^{i-1})^2(1+\phi)C_3^2\gamma^{2(i-2)}\beta_i,\quad LA+1\leq i\leq L-1 \tag{12}$$

$$\lambda_3 = 4\beta_iC_3^2\gamma^{2(i-2)}\theta(1+3L_1^{-1}L_3\gamma^{i-1})^2,\ 2\leq i\leq LA \tag{13}$$

$$\lambda_4 = \frac{1}{6}\beta_iC_3^2\gamma^{2(i-2)}(1+3L_1^{-1}L_3\gamma^{i-1})^2(1+(1-\theta)\phi),\quad LA+1\leq i\leq L-1 \tag{14}$$

$$\lambda_5 = L_1\beta_1\|X\|_F(4C_3 + L_0\sqrt{nm_l}\gamma^{-1}(1 + \sqrt{\frac{2L_3C_3\|X\|_F\gamma}{L_1(4C_3 + L_0\sqrt{nm_i})}}). \tag{15}$$

With these defined constants, we impose some conditions on the penalty parameters in the augmented Lagrangian, the regularization parameter λ, and the initializations as follows:

$$\beta_L \geq 3.5, \tag{16}$$

$$\frac{\beta_{L-1}}{\beta_L} \geq \max\{7\gamma^2,\ \frac{38\gamma^2L_1^2\alpha^{(L-LA)}}{3(3L_1^2\alpha_1^{(L-LA)} - 32)}\}, \tag{17}$$

$$\frac{\beta_i}{\beta_{i+1}} \geq \max\{36\sqrt{L}(2L_1^2 + (4L_3+L_2)C_3\gamma^i)\gamma^2,\ 6(\sqrt{3L_1} + \sqrt{2L_3C_3\gamma^i})^2\gamma^2,\ i = LA+1, ..., L-2 \tag{18}$$

$$\frac{(\beta')_i}{\beta_{i+1}} \geq \max\{6\sqrt{L}(2L_1^2 + (4L_3+L_2)C_3\gamma^i)\gamma^2,\ \sqrt{12 + 64LA[2L_1^2 + (4L_3+L_2)C_3\gamma]^2\gamma^4},\ 6(\sqrt{3L_1} + \sqrt{2L_3C_3\gamma^i})^2\gamma^2,\ \frac{\alpha_1+\alpha_2}{2+2\phi+\sqrt{24L+16\phi L}},\ \frac{(\alpha_1+\alpha_2)\gamma^2}{34\sqrt{LA}}\},\ i = 1, ..., LA \tag{19}$$

$$\frac{\beta_i}{(\beta')_i} \geq \max\{7,\ \frac{24L_1 - (1+3LA)L_1^2 + 1}{12(24L+1-L_1^2)},\ \frac{(\beta')^2}{\beta^2}\},\ i = 1, ..., LA \tag{20}$$

$$\frac{(\beta')_{LA}}{\beta_{L-1}} \geq 6(L_1\gamma)^{L-LA-1}, \tag{21}$$

$$\alpha_1 + \alpha_2 \geq \frac{3}{8L}, \tag{22}$$

$$\lambda \geq \max\{12\beta_LC_3^2\gamma^{2L-4},\ \frac{1}{5} + \frac{4\phi}{3L_1^2},\ \lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4,\ \lambda_5\}, \tag{23}$$

$$m_{\min} \geq \max\{\frac{\sqrt{24L+1}L_1 - \sqrt{18L_1}}{n(24L_0L_3)^2},\ \frac{\sqrt{96LA+1}L_1 - \sqrt{18L_1}}{n(24L_0L_3)^2},\ \frac{\sqrt{24L+16\phi L+1}L_1 - \sqrt{18L_1}}{n(24L_0L_3)^2}, 0\}, \tag{24}$$

$$\|W_0^i\|_F \leq \gamma,\ \|V_0^i\|_F \leq 3C_3\gamma^{i-1},$$

$$\|U\|_F \leq 3C_3\beta_i\gamma^{i-1},\ i = 1, ..., L \tag{25}$$

$$\|(U')\|_F \leq 3C_3(\beta')_i\gamma^{i-1},\ i = 1, ..., LA \tag{26}$$

Under these assumptions, we state the main convergence theorem of ADMM-CNN as follows.

B. Dual Expressed by Primal

In the following lemma,we we will show that the updates of dual variable $\{U_i^k\}_{i=1^L}$ can be expressed explicitly by the updates of primal variables.

Lemma 1: Suppose that all assumptions hold. And the$\{\mathcal{Q}^k = (\{W_i^k\}_{i=1}^L,\ \{V_i^k\}_{i=1}^L,\ \{d_i^k\}_{i=1}^{LA},\ \{U_i^k\}_{i=1}^L),$

$\{(U')_i^k\}_{i=1}^{LA})\}$ is the sequence generated by the ADMM-CNN algorithm.

$$U_L^k = V_L^k - Y, \tag{27}$$

$$\begin{aligned} U_{L-1}^k &= (W_L^k)^T[U_L^{k-1} + \beta_L(W_L^k V_{L-1}^k - V_L^{k-1})] \\ &= (W_L^k)^T[U_L^{k-1} + \beta_L(V_L^k - V_L^{k-1})], \end{aligned} \tag{28}$$

$$\begin{aligned} U_l^k &= (W_{l+1}^k)^T[(U_{l+1}^{k-1} + \beta_L(\sigma(W_{l+1}^k V_l^{k-1}) - V_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1})] + \frac{\beta_{l+1} t_l^k}{2}(W_{l+1}^k)^T W_{l+1}^k (V_l^k - V_l^{k-1}) \\ &= (W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\ &\quad + \beta'(W_{l+1}^k)^T[((\sigma(W_{l+1}^k V_l^{k-1} \\ &\quad - \sigma(W_{l+1}^k V_l^k)) + (V_{l+1}^k - V_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1}) + t_l^k W_{l+1}^k (V_l^k - V_l^{k-1})/2], \end{aligned} \tag{29}$$

$$U_l^k = (\beta')_l(V_l^k - V_l^{k-1}) + (U')_l^k, \tag{30}$$

$$\begin{aligned} (U')_l^k &= (W_{l+1}^k)^T[(U_{l+1}^{k-1} + \beta_L(\sigma(W_{l+1}^k V_l^{k-1}) - d_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1})] + \frac{\beta_{l+1} t_l^k}{2}(W_{l+1}^k)^T W_{l+1}^k (V_l^k - V_l^{k-1}) \\ &= (W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\ &\quad + \beta'(W_{l+1}^k)^T[((\sigma(W_{l+1}^k V_l^{k-1} \\ &\quad - \sigma(W_{l+1}^k V_l^k)) + (d_{l+1}^k - d_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1}) + t_l^k W_{l+1}^k (V_l^k - V_l^{k-1})/2]. \end{aligned} \tag{31}$$

Proof: These conclusions can be reached based on the explicit updates of primal parameters.

(1)According to the update of d_l, we can get

$$\begin{aligned} &\beta_l(\sigma(W_l^k * V_{l-1}^k) - d_l^k + \frac{U_l^{k-1}}{\beta_l}) \\ &+ (\beta')_l(\text{pool}(d_l^k) - V_l^{k-1} + \frac{(U')_l^{k-1}}{(\beta')_l}) = 0, \end{aligned} \tag{32}$$

and together with the update of U_l^k and $(U')_l^k$ in the algorithm yileds

$$U_l^k = (\beta')_l(V_l^k - V_l^{k-1}) + (U')_l^k. \tag{33}$$

It is worth noting that, to facilitate calculations, we have unified the matrices in the above equation to the same dimension by performing max pooling on the larger dimensions to obtain data of the same dimension.

(2)By the update of V_l^k, for $l = 1, ..., LA$,

$$\begin{aligned} &- [(U')_l^{k-1} + (\beta')(\text{pool}(d_l^k) - V_l^k)] \\ &+ \frac{\beta_{l+1} t_l^k}{2}(W_{l+1}^k)^T W_{l+1}^k (V_l^k - V_l^{k-1}) \\ &+ (W_{l+1}^k)^T[(U_{l+1}^{k-1} + \beta_{l+1}(\sigma(W_{l+1}^k V_l^{k-1}) - d_{l+1}^{k-1})) \\ &\odot \sigma'(W_{l+1}^k V_l^{k-1})] = 0, \end{aligned} \tag{34}$$

and together with the update of $(U')_l^k$ and U_{l+1}^k in the algorithm yields

$$\begin{aligned} (U')_l^k &= (W_{l+1}^k)^T[(U_{l+1}^{k-1} + \beta_L(\sigma(W_{l+1}^k V_l^{k-1}) - d_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1})] + \frac{\beta_{l+1} t_l^k}{2}(W_{l+1}^k)^T W_{l+1}^k (V_l^k - V_l^{k-1}) \\ &= (W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\ &\quad + \beta'(W_{l+1}^k)^T[((\sigma(W_{l+1}^k V_l^{k-1} \\ &\quad - \sigma(W_{l+1}^k V_l^k)) + (d_{l+1}^k - d_{l+1}^{k-1})) \\ &\quad \odot \sigma'(W_{l+1}^k V_l^{k-1}) + t_l^k W_{l+1}^k (V_l^k - V_l^{k-1})/2]. \end{aligned} \tag{35}$$

We have expressed the update of dual variables in detail using the update of primal variables and complete the proof of this lemma. ■

C. Proofs for Convergence

1) One-step Progress Lemma: The following lemma(One-step progress) play a key role for proving the sufficient descent lemma by presenting the progress made by a single Update of ADMM-CNN.

Lemma 2: Under the above assumptions and initializations, for any $k \leq 1$, there holds

$$\begin{aligned} &L(Q^k) \leq L(Q^{k-1}) \\ &- \sum_{l=1}^{L}(\frac{\lambda}{2}\|W_l^k - W_l^{k-1}\|_F^2 + \frac{\beta_l h_l^k}{4}\|(W_l^k - W_l^{k-1})V_{l-1}^{k-1}\|_F^2) \\ &- \sum_{l=LA+1}^{L-1}(\frac{\beta_l}{2}\|V_l^k - V_l^{k-1}\|_F^2 + \frac{\beta_{l+1} t_l^k}{4}\|W_{l+1}^k(V_l^k - V_l^{k-1})\|_F^2) \\ &- \sum_{l=1}^{LA}(\frac{(\beta')_l}{2}\|V_l^k - V_l^{k-1}\|_F^2 + \frac{\beta_{l+1} t_l^k}{4}\|W_{l+1}^k(V_l^k - V_l^{k-1})\|_F^2) \\ &- \frac{1+\beta_L}{2}\|V_L^k - V_L^{k-1}\|_F^2 \\ &- \sum_{l=1}^{LA}\frac{\beta_l + (\beta')_l}{2}\|\text{pool}(d_l^k) - \text{pool}(d_l^{k-1})\|_F^2 \\ &+ \sum_{l=1}^{L}\beta_l^{-1}\|U_l^k - U_l^{k-1}\|_F^2 + \sum_{l=1}^{LA}\|(U')_l^k - (U')_l^{k-1}\|_F^2. \end{aligned} \tag{36}$$

Before proving Lemma 2, we firstly give two lemma as follows,

Lemma 3: For a constant $c \in \mathbb{R}$, let f_c be the function on $\mathbb{R}$ given by $f_c(u) = (\sigma(u) - c)^2$. Then the following hold

$$f_c(v) \leq f_c(u) + f_c'(u)(v-u) + \frac{\mathbb{L}|c|}{2}(v-u)^2, \ \forall\ u,\ v \in \mathbb{R} \tag{37}$$

where $\mathbb{L}(\|c\|)$ are defined before.

Proof: It is easy to show that $|f_c''(u) \leq \mathbb{L}(|c|)$, $\forall u \in \mathbb{R}$, which yields the above inequality. ■

Lemma 4: Let variables be updated according to ADMM-

CNN algorithm, then we have

$$\begin{aligned}&\frac{\lambda}{2}\|W^k\|_F^2+\beta H_\sigma(W^k;A,B)\\&\le\frac{\lambda}{2}\|W^{k-1}\|_F^2+\beta H_\sigma(W^{k-1};A,B)\\&-\frac{\lambda}{2}\|W^k-W^{k-1}\|_F^2-\frac{\beta h^k}{4}\|(W^k-W^{k-1})A\|_F^2,\end{aligned}\tag{38}$$

$$\begin{aligned}&\frac{\lambda}{2}\|V^k-C\|_F^2+\beta M_\sigma(V^k;A,B)\\&\le\frac{\lambda}{2}\|V^k-C\|_F^2+\beta H_\sigma(V^k-C^{k-1};A,B)\\&-\frac{\lambda}{2}\|V^k-V^{k-1}\|_F^2-\frac{\beta t^k}{4}\|A(V^k-V^{k-1})\|_F^2.\end{aligned}\tag{39}$$

Proof: Let $h(W)=\frac{\lambda}{2}\|W\|^2+\beta H_\sigma^k(W;A,B)$ By applying optimality condition of updating W^k and Taylor's formula,there holds,

$$\begin{aligned}&h(W^{k-1})\\&=h(W^k)+\frac{\lambda}{2}\|W^k-W^{k-1}\|_F^2+\frac{\beta h^k}{4}\|(W^k-W^{k-1})A\|_F^2,\end{aligned}\tag{40}$$

which implies,

$$\begin{aligned}&\frac{\lambda}{2}\|W^{k-1}\|_F^2+\beta H_\sigma(W^{k-1};A,B)\\&=\frac{\lambda}{2}\|W^k\|_F^2+\beta(H_\sigma(W^{k-1};A,B)\\&+<\nabla H_\sigma(W^{k-1};A,B),W^k-W^{k-1}>\\&+\frac{h^k}{4}\|(W^k-W^{k-1})A\|_F^2)\\&+\frac{\lambda}{2}\|W^k-W^{k-1}\|_F^2-\frac{\beta h^k}{4}\|(W^k-W^{k-1})A\|_F^2\\&\ge\frac{\lambda}{2}\|W^k\|_F^2+\beta H_\sigma(W^k;A,B)\\&+\frac{\lambda}{2}\|W^k-W^{k-1}\|_F^2+\frac{\beta h^k}{4}\|(W^k-W^{k-1})A\|_F^2.\end{aligned}\tag{41}$$

■

Base on above lemmas, we prove Lemma 2 in the following,

Proof: Essentially, all inequalities can be obtained by applying optimality conditions of updating all variables respectively. We only prove the inequalities in convolution layers for convenience and the other inequalities follow the routines of these inequalities.The process is as follow.

(1)According to the update of W_l^k and lemma 4,we have:

$$\begin{aligned}&L(\overline{W}_l^{k-1},\{V_l^{k-1}\}_{l=1}^L,\{d_l^{k-1}\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&\le L(\overline{W}_l^{k},\{V_l^{k-1}\}_{l=1}^L,\{d_l^{k-1}\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&-\frac{\lambda}{2}\|W_l^k-W_l^{k-1}\|_F^2-\frac{\beta_l h_l^k}{4}\|(W_l^k-W_l^{k-1})V_{l-1}^{k-1}\|_F^2,\end{aligned}\tag{42}$$

where $\overline{W}_l^{k-1}=\{W_{<l}^{k-1},W_l^k,W_{>l}^k\}$ and $\overline{W}_l^k=\{W_{<l}^{k-1},W_l^{k-1},W_{>l}^k\}$.

(2)According to the update of V_l^k and lemma 4, we have:

$$\begin{aligned}&L(\{W_l^k\}_{l=1}^L,\overline{V}_l^{k-1},\{d_l^k\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&\le L(\{W_l^k\}_{l=1}^L,\overline{V}_l^{k},\{d_l^k\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&-\frac{\beta'}{2}\|V_l^k-V_l^{k-1}\|_F^2-\frac{\beta_{l+1}t_l^k}{4}\|W_{l+1}^k(V_l^k-V_l^{k-1})\|_F^2,\end{aligned}\tag{43}$$

where $\overline{V}_l^{k-1}=\{V_{<l}^{k-1},V_l^k,V_{>l}^k\}$ and $\overline{V}_l^k=\{V_{<l}^{k-1},V_l^{k-1},V_{>l}^k\}$.

(3)According to the update of d_l^k,:

$$\begin{aligned}&L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,\overline{d}_l^{k-1},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&\le L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,\overline{d}_l^{k-1},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&-\frac{\beta}{2}\|d_l^k-d_l^{k-1}\|_F^2-\frac{\beta_l'}{2}\|\text{pool}(d_l^k)-\text{pool}(d_l^{k-1})\|_F^2\\&=L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,d_{<l}^k,d_l^{k-1},d_{>l}^{k-1},\{U_l^{k-1}\}_{l=1}^L)\\&-\frac{\beta_l'+\beta_l}{2}\|\text{pool}(d_l^k)-\text{pool}(d_l^{k-1})\|_F^2.\end{aligned}\tag{44}$$

where $\overline{d}_l^{k-1}=\{d_{<l}^{k-1},d_l^k,d_{>l}^k\}$ and $\overline{d}_l^k=\{d_{<l}^{k-1},d_l^{k-1},d_{>l}^k\}$.

(4)The update of dual variables can be directly derived by subtracting the detailed Lagrangian functions before and after updating the dual parameters, and then substituting the update formula of dual variables into it.

$$\begin{aligned}&L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,\{d_l^k\}_{l=1}^{LA},\{U_l^k\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&=L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,\{d_l^k\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L,\{(U')_l^{k-1}\}_{l=1}^{LA})\\&+\sum_{l=1}^{LA}<U_l^k-U_l^{k-1},\ \sigma(W_l^k*V_{l-1}^k)-d_l^k>\\&+\sum_{l=1}^{LA}<(U')_l^k-(U')_l^{k-1},\ \text{pool}(d_l^k)-V_l^k>\\&+\sum_{l=1}^{L-1}<U_l^k-U_l^{k-1},\ \sigma(W_l^kV_{l-1}^k)-V_l^k>\\&+<U_L^k-U_L^{k-1},\ W_L^kV_{L-1}^k-V_L^k>\\&=L(\{W_l^k\}_{l=1}^L,\{V_l^k\}_{l=1}^L,\{d_l^k\}_{l=1}^{LA},\{U_l^{k-1}\}_{l=1}^L)\\&+\sum_{l=1}^{L}\beta^{-1}\|U_l^k-U_l^{k-1}\|_F^2+\sum_{l=1}^{LA}(\beta')^{-1}\|(U')_l^k-(U')_l^{k-1}\|_F^2.\end{aligned}\tag{45}$$

Summing up all the equations and inequations yields (36). ■

2) Dual-bounded-by-primal Lemma: From Lemma 2, there are two key parts that contribute to the progress along the augmented Lagrangian sequence, namely, the descent part arisen by the primal updates and the ascent part brought by the dual updates. Due to the existence of the dual ascent part, the convergence of nonconvex ADMM is usually very challenging. In order to further estimate the progress in terms of the primal updates, we shall bound these dual ascent parts via the primal updates. The following lemma shows that the dual ascent quantity can be bounded by the discrepancies between two successive primal updates.

Lemma 5: For any positive inter $k \leq 2$, the following hold

$$\|U_L^k - U_L^{k-1}\|_F = \|V_L^k - V_L^{k-1}\|_F, \tag{46}$$

$$\begin{aligned}
&\|U_{L-1}^k - U_{L-1}^{k-1}\|_F \\
&\leq \|W_L^k\|_F\|U_L^k - U_L^{k-1}\|_F + \|U_L^{k-1}\|_F\|W_L^k - W_L^{k-1}\|_F \\
&+ \beta_L\|W_L^k\|_F\|V_L^k - V_L^{k-1}\|_F \\
&+ \beta_L\|W_L^{k-1}\|_F\|V_L^{k-1} - V_L^{k-2}\|_F,
\end{aligned} \tag{47}$$

$$\begin{aligned}
&\|U_l^k - U_l^{k-1}\|_F \\
&\leq L_1\|W_{l+1}^k\|_F\|U_{l+1}^k - U_{l+1}^{k-1}\|_F + (L_1\|U_{l+1}^{k-1}\|_F \\
&+ L_2\|W_{l+1}^{k-1}\|_F\|U_{l+1}^{k-1}\|_F\|V_l^{k-1}\|_F)\|W_{l+1}^k - W_{l+1}^{k-1}\|_F \\
&+ L_2\beta_{l+1}(\|W_{l+1}^k\|_F\|V_{l+1}^k - V_{l+1}^{k-1}\|_F \\
&+ \|W_{l+1}^{k-1}\|_F\|V_{l+1}^{k-1} - V_{l+1}^{k-2}\|_F) \\
&+ (L_1^2 + \frac{t_l^k}{2})\beta_{l+1}\|W_{l+1}^k\|_F^2\|V_l^k - V_l^{k-1}\|_F \\
&+ ((L_1^2 + \frac{t_l^{k-1}}{2})\beta_{l+1} \\
&+ L_2\|U_{l+1}^{k-1}\|_F)\|W_{l+1}^{k-1}\|_F^2\|V_l^{k-1} - V_l^{k-2}\|_F,
\end{aligned} \tag{48}$$

$$\begin{aligned}
&\|U_l^k - U_l^{k-1}\|_F \\
&\leq (\beta')_l(\|V_l^k - V_l^{k-1}\|_F \\
&+ \|V_l^{k-1} - V_l^{k-2}\|_F) + \|(U')_l^k - (U')_l^{k-1}\|_F,
\end{aligned} \tag{49}$$

$$\begin{aligned}
&\|(U')_l^k - (U')_l^{k-1}\|_F \\
&\leq L_1\|W_{l+1}^k\|_F\|U_{l+1}^k - U_{l+1}^{k-1}\|_F + (L_1\|U_{l+1}^{k-1}\|_F \\
&+ L_2\|W_{l+1}^{k-1}\|_F\|U_{l+1}^{k-1}\|_F\|V_l^{k-1}\|_F)\|W_{l+1}^k - W_{l+1}^{k-1}\|_F \\
&L_1\beta_{l+1}(\|W_{l+1}^k\|_F\|d_{l+1}^k - d_{l+1}^{k-1}\|_F \\
&+ \|W_{l+1}^{k-1}\|_F\|d_{l+1}^{k-1} - d_{l+1}^{k-2}\|_F) \\
&(L_1^2 + \frac{t_l^k}{2})\beta_{l+1}\|W_{l+1}^k\|_F^2\|V_l^k - V_l^{k-1}\|_F \\
&((L_1^2 + \frac{t_l^{k-1}}{2})\beta_{l+1} \\
&+ L_2\|U_{l+1}^{k-1}\|_F)\|W_{l+1}^{k-1}\|_F^2\|V_l^{k-1} - V_l^{k-2}\|_F.
\end{aligned} \tag{50}$$

Proof: (1) By the update of U_l^k and the triangle inequality, there holds

$$\begin{aligned}
&U_l^k - U_l^{k-1} \\
&= (\beta')_l(V_l^k - V_l^{k-1}) + (U')_l^k \\
&- (\beta')_l(V_l^{k-1} - V_l^{k-2}) - (U')_l^{k-1} \\
&= (\beta')_l[(V_l^k - V_l^{k-1}) + (V_l^{k-1} - V_l^{k-2})] \\
&+ (U')_l^{k-1} - (U')_l^k \\
&\leq (\beta')_l(\|V_l^k - V_l^{k-1}\|_F + \|V_l^{k-1} - V_l^{k-2}\|_F) \\
&+ \|(U')_l^{k-1} - (U')_l^k\|_F.
\end{aligned} \tag{51}$$

(2)By the update of $(U')_l^k$, there holds

$$\begin{aligned}
&(U')_l^k - (U')_l^{k-1} \\
&= (W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\
&- (W_{l+1}^{k-1})^T(U_{l+1}^{k-1} \odot \sigma'(W_{l+1}^{k-1} V_l^{k-2})) \\
&+ \beta_{l+1}(W_{l+1}^k)^T[((\sigma(W_{l+1}^k V_l^{k-1}) - \sigma(W_{l+1}^k V_l^k)) \\
&+ (d_{l+1}^k - d_{l+1}^{k-1})) \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\
&+ \frac{t_l^k}{2}W_{l+1}^k(V_l^k - V_l^{k-1})] \\
&- \beta_{l+1}(W_{l+1}^{k-1})^T[((\sigma(W_{l+1}^{k-1} V_l^{k-2}) - \sigma(W_{l+1}^{k-1} V_l^{k-1})) \\
&+ (d_{l+1}^{k-1} - d_{l+1}^{k-2})) \odot \sigma'(W_{l+1}^{k-1} V_l^{k-2})) \\
&+ \frac{t_l^{k-1}}{2}W_{l+1}^{k-1}(V_l^{k-1} - V_l^{k-2})].
\end{aligned} \tag{52}$$

By triangle inequality, the above inequality implies,

$$\begin{aligned}
&\|(U')_l^k - (U')_l^{k-1}\| \\
&= \|(W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\
&- (W_{l+1}^{k-1})^T(U_{l+1}^{k-1} \odot \sigma'(W_{l+1}^{k-1} V_l^{k-2}))\| \\
&+ \beta_{l+1}\|W_{l+1}^k\|((c^2 + \frac{t_l^k}{2}W^k)\|(W_{l+1}^k\|_F\|V_l^k - V_l^{k-1}\|_F \\
&+ L_1\|d_{l+1}^k - d_{l+1}^{k-1}\|_F \\
&+ \beta_{l+1}\|W_{l+1}^{k-1}\|((c^2 + \frac{t_l^{k-1}}{2}W^k)\|(W_{l+1}^{k-1}\|_F\|V_l^{k-1} - V_l^{k-2}\|_F \\
&+ L_1\|d_{l+1}^{k-1} - d_{l+1}^{k-2}\|_F.
\end{aligned} \tag{53}$$

Similarly,there holds

$$\begin{aligned}
&\|(W_{l+1}^k)^T(U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})) \\
&- (W_{l+1}^{k-1})^T(U_{l+1}^{k-1} \odot \sigma'(W_{l+1}^{k-1} V_l^{k-2}))\| \\
&\leq \|W_{l+1}^k - W_{l+1}^{k-1}\|_F\|U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})\|_F \\
&+ \|W_{l+1}^{k-1}\|_F\|U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1}) \\
&- U_{l+1}^{k-1} \odot \sigma'(W_{l+1}^{k-1} V_l^{k-2})\|_F \\
&\leq \|W_{l+1}^k - W_{l+1}^{k-1}\|_F\|U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})\|_F \\
&+ \|W_{l+1}^{k-1}\|_F(\|(U_{l+1}^k - U_{l+1}^{k-1}) \odot \sigma'(W_{l+1}^k V_l^{k-1})\|_F \\
&+ \|U_{l+1}^{k-1} \odot (\sigma'(W_{l+1}^k V_l^{k-1}) - \sigma'(W_{l+1}^{k-1} V_l^{k-2}))\|_F) \\
&\leq \|W_{l+1}^k - W_{l+1}^{k-1}\|_F\|U_{l+1}^k \odot \sigma'(W_{l+1}^k V_l^{k-1})\|_F \\
&+ \|W_{l+1}^{k-1}\|_F(L_1\|(U_{l+1}^k - U_{l+1}^{k-1})\|_F \\
&+ L_2\|U_{l+1}^{k-1}\|_F\|W_{l+1}^k V_l^{k-1} - W_{l+1}^{k-1} V_l^{k-2}\|_F) \\
&\leq L_1\|U_{l+1}^k\|_F\|W_{l+1}^k - W_{l+1}^{k-1}\|_F \\
&+ \|W_{l+1}^{k-1}\|_F(L_1\|(U_{l+1}^k - U_{l+1}^{k-1})\|_F \\
&+ L_2\|U_{l+1}^{k-1}\|_F(\|W_{l+1}^k - W_{l+1}^{k-1}\|_F\|V_l^{k-1}\|_F \\
&+ \|W_{l+1}^{k-1}\|_F\|V_l^{k-1} - V_l^{k-2}\|_F)).
\end{aligned} \tag{54}$$

This completes the proof of this lemma. ■

3) Boundedness Lemma: We hope that the upper bounds of the dual variables are only related to the expected terms. However, by observing the above upper bounds, we can see that the desired terms are multiplied by lot of other terms. As a result, we require some other bound property of the sequence.

Lemma 6: Under the assumption, there hold

$$\|W_i^k\|_F \le \gamma,\ \|V_i^k\|_F \le 3C_3\gamma^{i-1},$$
$$\|U_i^k\|_F \le C_3\beta_i\gamma^{i-1},\ i=1,...,L \tag{55}$$
$$\|d_i^k\|_F \le 3C_3\gamma^{i-1},$$
$$\|(U')_i^k\|_F \le C_3\gamma^{i-1},\ i=1,...,LA \tag{56}$$
$$h_i^k \le 4L_3C_3\gamma^{i-1}, \tag{57}$$
$$t_i^k \le 4L_3C_3\gamma^{i}. \tag{58}$$

Proof: For convenience,we only prove the boundedness condition for $k=1$.

(1)By the update of W_l^1

$$\|W_l^1\|_F \le (1-\frac{\lambda}{\lambda+18\beta_lL_3C_3^3\gamma^{3l-5}})\gamma + \frac{3L_1\beta_lC_3\gamma^{l-2}(4C_3\gamma^{l-1}+L_0\sqrt{nm_l})}{\lambda}. \tag{59}$$

To make $\|W_l^1\|_F \le \gamma$, we can inversely solve the above equation, and we obtain:

$$\lambda \ge \frac{a_l+\sqrt{a_l^2+4a_lb_l}}{2}, \tag{60}$$

where $a_l = 3L_1\beta_lC_3\gamma^{l-3}(4C_3\gamma^{l-1}+L_0\sqrt{nm_l})$ and $b_l = 18\beta_lL_3C_3^3\gamma^{3l-5}$. And by the inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, there holds

$$\lambda \ge a_l(1+\sqrt{\frac{b_l}{a_l}}) \ge \frac{a_l+\sqrt{a_l^2+4a_lb_l}}{2}, \tag{61}$$

which indicates the range or value of λ

$$\lambda \ge 3L_1\beta_lC_3\gamma^{l-3}(4C_3\gamma^{l-1}+L_0\sqrt{nm_l})(1 + \sqrt{\frac{6\beta_lL_3C_3^3\gamma^{3l-5}}{L_1\beta_lC_3\gamma^{l-3}(4C_3\gamma^{l-1}+L_0\sqrt{nm_l})}}). \tag{62}$$

(2)By the update of d_l^1,

$$\|d\|_F \le \frac{\beta_lL_0\sqrt{nm_l}+C_3\beta_l\gamma^{l-1}+4(\beta')_lC_3\gamma^{l-1}}{\beta_l+(\beta')_l}. \tag{63}$$

To make $\|d_l^1\|_F \le 3C_3\gamma^{l-1}$, there holds

$$\frac{\frac{\beta_l}{(\beta')_l}(L_0\sqrt{nm_l}+C_3\gamma^{l-1})+4C_3\gamma^{l-1}}{1+\frac{\beta_l}{(\beta')_l}} \le 3C_3\gamma^{l-1}, \tag{64}$$

which implies

$$\frac{\beta_l}{(\beta')_l} \ge \frac{C_3\gamma^{l-1}}{2C_3\gamma^{l-1}-L_0\sqrt{nm_l}}. \tag{65}$$

(3)By the update of V_l^1,

$$\|V_l^1\|_F \le (1-\frac{\rho_l}{\rho_l+2L_3C_3\gamma^{l+2}})\cdot 3C_3\gamma^{l-1} + (C_3\gamma^{l-1}+L_0\sqrt{nm_l}) + \frac{L_1\gamma(4C_3\gamma^{l}+L_0\sqrt{nm_{l+1}})}{\rho_l}, \tag{66}$$

where $\rho=\frac{(\beta')_l}{\beta_{l+1}}$. To guarantee $\|V_l^1\|_F \le 3C_3\gamma^{l-1}$, we have

$$\rho_l \ge \frac{\bar{b}_l+\sqrt{\bar{b}_l^2+4\bar{a}_l\bar{c}_l}}{2\bar{a}_j},$$

where $\bar{a}_l = 2-\frac{L_0\sqrt{nm_l}}{C_3\gamma^{l-1}}$, $\bar{b}_l = 2L_3C_3\gamma^{l+2}+2L_3C_3\gamma^3L_0\sqrt{nm_l}+4L_3\gamma^2+\frac{L_1L_0\sqrt{nm_{l+1}}}{C_3\gamma^{l-2}}$, and $\bar{c}_l = 2L_1L_3\gamma^4(4C_3\gamma^l+L_0\sqrt{nm_l})$. And we can get that,

$$\bar{a}_l \ge \frac{3}{2},\ \bar{b}_l \le (4.5L_1+3L_3C_3\gamma^l)\gamma^2,\ \bar{c}_l \le 9L_1L_3C_3\gamma^{l+4}$$

Thus, the following holds

$$\frac{\bar{b}_l+\sqrt{\bar{b}_l^2+4\bar{a}_l\bar{c}_l}}{2\bar{a}_l} \le \frac{1}{3}\bar{b}_l(1+\sqrt{1+\frac{6\bar{c}_l}{\bar{b}_l^2}}) \le \frac{2}{3}\bar{b}_l+\frac{\sqrt{6\bar{c}_l}}{3} \le (3L_1+\sqrt{6L_1L_3C_3\gamma^j}+2L_3C_3\gamma^l)\gamma^2, \tag{67}$$

which implies that the penalty parameter should also be greater than this value.

(4)By the update of U_l^1,

$$\begin{aligned}\|U_l^k\| &= (\beta')_l(V_l^k-V_l^{k-1})+(U')_l^{k-1}\\ &\le (\beta')_l(3C_3\gamma^{l-1}+3C_3\gamma^{l-1})+(\beta')_lC_3\gamma^{l-1}\\ &= 7(\beta')_lC_3\gamma^{l-1}.\end{aligned}$$

To make $U_l^k \le \beta_lC_3\gamma^{l-1}$, it requires

$$\frac{\beta_l}{(\beta')_l} \ge 7. \tag{68}$$

By the update $(U')_l^1$,

$$\begin{aligned}&\|(U')_l^k\|\\ &\le \beta_{l+1}\gamma(2L_1L_2\sqrt{nm_l+1}+7L_1C_3\gamma^l+12L_3C_3^2\gamma^{2l})\\ &\le C_3\beta_{l+1}\gamma^{l+1}(8L_1+12L_3C_3^2\gamma^l)\\ &\le C_3^2(\beta')_l\gamma^{l-1},\end{aligned}$$

which means

$$\frac{(\beta')_l}{\beta_{l+1}} \ge 6(\sqrt{3L_1}+\sqrt{2L_3C_3\gamma^l})^2\gamma^2 \ge (8L_1+12L_3C_3^2\gamma^l)\gamma^2. \tag{69}$$

Therefore, we have shown that (55)-(58) hold for $k=1$. Similarly, we can show that once (55)-(58) hold for some k, then they will hold for $k+1$. Hence, we can show (55)-(58) hold for any $k\in N$ recursively. This completes the proof of this lemma. ■

4) Sufficient Descent Lemma:

Lemma 7: To prove sufficient descent lemma, we first present a key lemma based on Lemma 5 and Lemma 6. Under the assumption, for any $k\ge 2$, there holds,

$$\|U_L^k-U_L^{k-1}\|_F = \|V_L^k-V_L^{k-1}\|_F, \tag{70}$$
$$\|U_{L-1}^k-U_{L-1}^{k-1}\|_F \le C_3\beta_L\gamma^{L-1}\|W_L^k-W_L^{k-1}\|_F + \gamma(1+\beta_L)\|V_L^k-V_L^{k-1}\|_F + \beta_L\gamma\|V_L^{k-1}-V_L^{k-2}\|_F, \tag{71}$$
$$\tag{72}$$

$$\|U_l^k - U_l^{k-1}\|_F \le \varepsilon_{1,l}^k + \varepsilon_{2,l}^k + \varepsilon_{3,l}^k,\ i = LA+1, \dots, L-2 \tag{73}$$

$$\|(U')_l^k - (U')_l^{k-1}\|_F \le \theta_{1,l}^k + \theta_{2,l}^k + \theta_{3,l}^k + \theta_l^k,\ i = 1, \dots, LA \tag{74}$$

$$\begin{aligned}\|U_l^k - U_l^{k-1}\|_F \le (\beta')_l(&\|V_l^k - V_l^{k-1}\|_F \\ &+ \|V_l^{k-1} - V_l^{k-2}\|_F) \\ &+ \|(U')_l^k - (U')_l^{k-1}\|_F,,\ i = 1, \dots, LA\end{aligned} \tag{75}$$

where

$$\begin{aligned}\varepsilon_{1,l}^k &= (L_1\gamma)^{L-l}L_1^{-1}C_3\beta_L\gamma^{L-2}\|W_L^k - W_L^{k-1}\|_F \\ &+ \sum_{i=l+1}^{L-1}(L_1\gamma)^{i-l}(C_3\beta_i\gamma^{i-2} \\ &+ 3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3})\|W_i^k - W_i^{k-1}\|_F,\end{aligned} \tag{76}$$

$$\begin{aligned}\varepsilon_{2,l}^k &= (L_1\gamma)^{L-l}L_1^{-1}(1+\beta_L)\|V_L^k - V_L^{k-1}\|_F \\ &+ (L_1\gamma)^{L-1-l}\beta_{L-1}\|V_{L-1}^k - V_{l-1}^{k-1}\|_F \\ &+ \sum_{i=l+1}^{L-2}(L_1\gamma)^{i-l}(\beta_i + (L_1^2 \\ &+ 2L_3C_3\gamma^i)\beta_{i+1}\gamma^2)\|V_i^k - V_i^{k-1}\|_F \\ &+ (L_1^2 + 2L_3C_3\gamma^l)\beta_{l+1}\gamma^2\|V_l^k - V_l^{k-1}\|_F,\end{aligned} \tag{77}$$

$$\begin{aligned}\varepsilon_{3,l}^k &= (L_1\gamma)^{L-l}\beta_L L_1^{-1}\|V_L^{k-1} - V_L^{k-2}\|_F \\ &+ (L_1\gamma)^{L-1-l}\beta_{L-1}\|V_{L-1}^{k-1} - V_{l-1}^{k-2}\|_F \\ &+ \sum_{i=l+1}^{L-2}(L_1\gamma)^{i-l}(\beta_i + (L_1^2 \\ &+ 2L_3C_3\gamma^i + L_2C_3\gamma^i)\beta_{i+1}\gamma^2)\|V_i^{k-1} - V_i^{k-2}\|_F \\ &+ (L_1^2 + 2L_3C_3\gamma^l + L_2C_3\gamma^l)\beta_{l+1}\gamma^2\|V_l^{k-1} - V_l^{k-2}\|_F,\end{aligned} \tag{78}$$

$$\begin{aligned}\theta_{1,l}^k &= \sum_{i=l+1}^{LA}(L_1\gamma)^{i-l}[(C_3\beta_i\gamma^{i-2} \\ &+ 3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3})\|W_i^k - W_i^{k-1}\|_F \\ &+ \beta_i(\|d_i^k - d_i^{k-1}\|_F + \|d_i^{k-1} - d_i^{k-2}\|_F)],\end{aligned} \tag{79}$$

$$\begin{aligned}\theta_{2,l}^k &= (L_1\gamma)^{LA-1}(\beta')_{LA}\|V_{LA}^k - V_{LA}^{k-1}\|_F \\ &+ \sum_{i=l+1}^{LA-l}(L_1\gamma)^{i-1}[(\beta')_i + (L_1^2 \\ &+ 2L_3C_3\gamma^i)\beta_{i+1}\gamma^2]\|V_i^k - V_i^{k-1}\|_F \\ &+ (L_1^2 + 2L_3C_3\gamma^j)\beta_{l+1}\gamma^2\|V_l^k - V_l^{k-1}\|_F,\end{aligned} \tag{80}$$

$$\begin{aligned}\theta_{3,l}^k &= (L_1\gamma)^{LA-1}(\beta')_{LA}\|V_{LA}^{k-1} - V_{LA}^{k-2}\|_F \\ &+ \sum_{i=l+1}^{LA-1}(L_1\gamma)^{i-l}[(\beta')_i + (L_1^2 \\ &+ 2L_3C_3\gamma^i + L_2C_3\gamma^i)\beta_{i+1}\gamma^2]\|V_i^{k-1} - V_i^{k-2}\|_F \\ &+ (L_1^2 + 2L_3C_3\gamma^l + L_2C_3\gamma^l)\beta_{l+1}\gamma^2\|V_l^{k-1} - V_l^{k-2}\|_F,\end{aligned} \tag{81}$$

$$\theta_l^k = (L_1\gamma)^{LA-l}\|(U')_{LA}^k - (U')_{LA}^{k-1}\|_F. \tag{82}$$

Moreover, the above inequalities imply for some constant $\alpha > 0$, there holds

$$\begin{aligned}&\sum_{l=1}^{L}\|U_l^k - U_l^{k-1}\|_F^2 + \sum_{l=1}^{LA}\|(U')_l^k - (U')_l^{k-1}\|_F \\ &\le \alpha(\sum_{l=1}^{L}(\|W_l^k - W_l^{k-1}\|_F^2 + \|V_l^k - V_l^{k-1}\|_F^2 \\ &+ \|V_l^{k-1} - V_l^{k-2}\|_F^2) \\ &+ \sum_{l=1}^{LA}(\|d_l^k - d_l^{k-1}\|_F^2 + \|d_l^{k-1} - d_l^{k-2}\|_F^2)).\end{aligned} \tag{83}$$

Proof: Similarly, we only prove the convolutional part. By Lemma 5 and Lemma 6,

$$\begin{aligned}&\|(U')_l^k - (U')_l^{k-1}\|_F \\ &\le L_1\gamma\|U_{l+1}^k - U_{l+1}^{k-1}\|_F + T_{l+1}^k + I_{l+1}^k,\end{aligned} \tag{84}$$

where

$$\begin{aligned}T_{l+1}^k &= (L_1C_3\beta_{l+1}\gamma^l + 3C_3^2L_2\beta_{l+1}\gamma^{2l})\|W_{l+1}^k - W_{l+1}^{k-1}\|_F \\ &+ L_1\gamma\beta_{l+1}(\|d_{l+1}^k - d_{l+1}^{k-1}\|_F + \|d_{l+1}^{k-1} - d_{l+1}^{k-2}\|_F),\end{aligned}$$

and

$$\begin{aligned}I_l^k &= (L_1^2 + 2L_3C_3\gamma^l)\beta_{l+1}\gamma^2\|V_l^k - V_l^{k-1}\|_F \\ &+ (L_1^2 + 2L_3C_3\gamma^l + L_2C_3\gamma^l)\beta_{l+1}\gamma^2\|V_l^{k-1} - V_l^{k-2}\|_F.\end{aligned}$$

By Lemma 5 and Lemma 6, for $l = 1, ..., LA$,

$$\begin{aligned}&\|U_l^k - U_l^{k-1}\|_F \\ &\le (\beta')_l(\|V_l^k - V_l^{k-1}\|_F + \|V_l^{k-1} - V_l^{k-2}\|_F) \\ &+ \|(U')_l^k - (U')_l^{k-1}\|_F.\end{aligned} \tag{85}$$

Let

$$\begin{aligned}M_{l+1}^k &= (L_1C_3\beta_{l+1}\gamma^l + 3C_3^2L_2\beta_{l+1}\gamma^{2l})\|W_{l+1}^k - W_{l+1}^{k-1}\|_F \\ &+ L_1\gamma[\beta_{l+1}(\|d_{l+1}^k - d_{l+1}^{k-1}\|_F + \|d_{l+1}^{k-1} - d_{l+1}^{k-2}\|_F) \\ &+ (\beta')_l(\|V_{l+1}^k - V_{l+1}^{k-1}\|_F + \|V_{l+1}^{k-1} - V_{l+1}^{k-2}\|_F),\end{aligned} \tag{86}$$

which means

$$\begin{aligned}&\|(U')_l^k - (U')_l^{k-1}\|_F \\ &\le L_1\gamma\|(U')_{l+1}^k - (U')_{l+1}^{k-1}\|_F + M_{l+1}^k + I_l^k.\end{aligned} \tag{87}$$

By the above inequality, we have

$$\begin{aligned}&\|(U')_l^k - (U')_l^{k-1}\|_F \\ &\le (L_1\gamma)^{LA-l}\|(U')_{LA}^k - (U')_{LA}^{k-1}\|_F + (L_1\gamma)^{LA-1-l}M_{LA}^k \\ &+ \sum_{i=1}^{LA-1-l}(L_1\gamma)^{i-1}(M_{l+i}^k + L_1\gamma I_{l+i}^k) + I_l^k.\end{aligned} \tag{88}$$

And By Lemma 5 and Lemma 6,

$$\begin{aligned}&\|(U')_{LA}^k - (U')_{LA}^{k-1}\|_F \\ &\le (L_1\gamma)^{L-1-LA}\|U_{L-1}^k - U_{L-1}^{k-1}\|_F + (L_1\gamma)^{L-2-LA}K_{L-1}^k \\ &+ \sum_{i=1}^{L-2-LA}(L_1\gamma)^{i-1}(K_{LA+i}^k + L_1\gamma P_{LA+i}^k) + P_{LA}^k,\end{aligned} \tag{89}$$

where

$$T_{j+1}^k = (L_1C_3\beta_{j+1}\gamma^j + 3C_3^2L_2\beta_{j+1}\gamma^{2j})\|W_{j+1}^k - W_{j+1}^{k-1}\|_F + L_1\gamma\beta_{j+1}(\|V_{j+1}^k - V_{j+1}^{k-1}\|_F + \|V_{j+1}^{k-1} - V_{j+1}^{k-2}\|_F),$$

and

$$P_j^k = (L_1^2 + 2L_3C_3\gamma^j)\beta_{j+1}\gamma^2\|V_j^k - V_j^{k-1}\|_F + (L_1^2 + 2L_3C_3\gamma^j + L_2C_3\gamma^j)\beta_{j+1}\gamma^2\|V_j^{k-1} - V_j^{k-2}\|_F. \tag{90}$$

Based on the aforementioned equations, we can derive that $\|(U')_{LA}^k - (U')_{LA}^{k-1}\|_F$ is bounded. Consequently, substitute the boundness and we can get,

$$\begin{aligned}
&\|(U')_l^k - (U')_l^{k-1}\|_F \\
&\le (L_1\gamma)^{LA-l}((L_1\gamma)^{L-1-LA}\|U_{L-1}^k - U_{L-1}^{k-1}\|_F \\
&+ (L_1\gamma)^{L-2-LA}K_{L-1}^k \\
&+ \sum_{i=1}^{L-2-LA}(L_1\gamma)^{i-1}(K_{LA+i}^k + L_1\gamma P_{LA+i}^k) + P_{LA}^k) \\
&+ (L_1\gamma)^{LA-1-l}T_{LA}^k + T_l^k \\
&+ \sum_{i=1}^{LA-1-l}(L_1\gamma)^{i-1}(M_{l+i}^k + L_1\gamma I_{l+i}^k).
\end{aligned} \tag{91}$$

Summing up all the above inequalities. By the inequality $(\sum_{i=1}^p u_i)^2 \le p\sum_{i=1}^p u_i^2$, we complete the proof. ∎

Based on Lemma 2, Lemma 6 and Lemma 7, we prove sufficient descent lemma as follows.

Proof: By lemma 7, we get,

$$\|U_L^k - U_L^{k-1}\|_F^2 = \|V_L^k - V_L^{k-1}\|_F^2. \tag{92}$$

Using the inequality$(\sum_{i=1}^3 a_i)^2 \le 3\sum_{i=1}^3 a_i^2$,

$$\begin{aligned}
\|U_{L-1}^k - U_{L-1}^{k-1}\|_F^2 \le\ & 3C_2^2\beta_L^2\gamma^{2(L-1)}\|W_L^k - W_L^{k-1}\|_F^2 \\
&+ 3\gamma^2(1+\beta_L)^2\|V_L^k - V_L^{k-1}\|_F^2 \\
&+ 3\beta_L^2\gamma^2\|V_L^{k-1} - V_L^{k-2}\|_F^2,
\end{aligned} \tag{93}$$

and for $l = LA+1, ..., L-2$, there holds,

$$\begin{aligned}
&\|U_l^k - U_l^{k-1}\|_F^2 \\
&\le 2(\varepsilon_{1,l}^k)^2 + 2(\varepsilon_{2,l}^k + \varepsilon_{3,l}^k)^2 \\
&\le 2(\varepsilon_{1,l}^k)^2 + 4((\varepsilon_{2,l}^k)^2 + (\varepsilon_{3,l}^k)^2)^2 \\
&\le 2(L-1)\Gamma_{1,l}^k + 4(L-l+1)(\Gamma_{2,l}^k + \Gamma_{3,l}^k),
\end{aligned} \tag{94}$$

where

$$\begin{aligned}
\Gamma_{1,l}^k =\ & (L_1\gamma)^{2(L-l)}L_1^{-2}C_3^2\beta_L^2\gamma^{2(L-2)}\|W_L^k - W_L^{k-1}\|_F^2 \\
&+ \sum_{i=l+1}^{L-1}(L_1\gamma)^{2(i-l)}(C_3\beta_i\gamma^{i-2} \\
&+ 3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3})^2\|W_i^k - W_i^{k-1}\|_F^2,
\end{aligned} \tag{95}$$

$$\begin{aligned}
\Gamma_{2,l}^k =\ & (L_1\gamma)^{2(L-l)}L_1^{-2}(1+\beta_L)^2\|V_L^k - V_L^{k-1}\|_F^2 \\
&+ (L_1\gamma)^{2(L-1-l)}\beta_{L-1}^2\|V_{L-1}^k - V_{l-1}^{k-1}\|_F^2 \\
&+ \sum_{i=l+1}^{L-2}(L_1\gamma)^{2(i-l)}[\beta_i + (L_1^2 \\
&+ 2L_3C_3\gamma^i)\beta_{i+1}\gamma^2]^2\|V_i^k - V_i^{k-1}\|_F^2 \\
&+ (L_1^2 + 2L_3C_3\gamma^l)^2\beta_{l+1}^2\gamma^4\|V_l^k - V_l^{k-1}\|_F^2,
\end{aligned} \tag{96}$$

$$\begin{aligned}
\Gamma_{3,l}^k =\ & (L_1\gamma)^{2(L-l)}L_1^{-2}\beta_L^2\|V_L^{k-1} - V_L^{k-2}\|_F^2 \\
&+ (L_1\gamma)^{2(L-1-l)}\beta_{L-1}^2\|V_{L-1}^{k-1} - V_{L-1}^{k-2}\|_F^2 \\
&+ \sum_{i=l+1}^{L-2}(L_1\gamma)^{2(i-l)}[\beta_i + (L_1^2 \\
&+ 2L_3C_3\gamma^i + L_2C_3\gamma^i)\beta_{i+1}\gamma^2]^2\|V_i^{k-1} - V_i^{k-2}\|_F^2 \\
&+ (L_1^2 + 2L_3C_3\gamma^l + L_2C_3\gamma^l)^2\beta_{l+1}^2\gamma^4\|V_l^{k-1} - V_l^{k-2}\|_F^2.
\end{aligned} \tag{97}$$

For the convolutional layers, there holds,

$$\begin{aligned}
&\|U_l^k - U_l^{k-1}\|_F^2 \\
&\le ((\beta')_l(\|V_l^k - V_l^{k-1}\|_F + \|V_l^{k-1} - V_l^{k-2}\|_F) \\
&+ \|(U')_l^k - (U')_l^{k-1}\|_F)^2 \\
&\le 3(\beta')_l^2\|V_l^k - V_l^{k-1}\|_F^2 + 3(\beta')_l^2\|V_l^{k-1} - V_l^{k-2}\|_F^2 \\
&+ 3\|(U')_l^k - (U')_l^{k-1}\|_F^2,
\end{aligned} \tag{98}$$

and

$$\begin{aligned}
&\|(U')_l^k - (U')_l^{k-1}\|_F^2 \\
&\le (\theta_{1,l}^k + \theta_{2,l}^k + \theta_{3,l}^k + \theta_l^k)^2 \\
&\le +4(\theta_{1,l}^k)^2 + 4(\theta_{2,l}^k)^2 + 4(\theta_{3,l}^k)^2 + 4(\theta_l^k)^2 \\
&\le 4(LA-l+2)H_{1,l}^k + 4(LA-l+1)H_{2,l}^k \\
&\quad + 4(LA-l+1)H_{3,l}^k + 4H_l^k,
\end{aligned} \tag{99}$$

where

$$\begin{aligned}
H_{1,l}^k =\ & \sum_{i=l+1}^{LA}(L_1\gamma)^{2(i-l)}[(C_3\beta_i\gamma^{i-2} \\
&+ 3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3})^2\|W_i^k - W_i^{k-1}\|_F^2 \\
&+ \beta_i^2(\|d_i^k - d_i^{k-1}\|_F^2 + \|d_i^{k-1} - d_i^{k-2}\|_F^2)],
\end{aligned} \tag{100}$$

$$\begin{aligned}
H_{2,l}^k =\ & (L_1\gamma)^{2(LA-1)}(\beta')_{LA}^2\|V_{LA}^k - V_{LA}^{k-1}\|_F^2 \\
&+ \sum_{i=l+1}^{LA-1}(L_1\gamma)^{2(i-l)}[(\beta')_i + (L_1^2 \\
&+ 2L_3C_3\gamma^i)\beta_{i+1}\gamma^2]^2\|V_i^k - V_i^{k-1}\|_F^2 \\
&+ (L_1^2 + 2L_3C_3\gamma^l)^2\beta_{l+1}^2\gamma^4\|V_l^k - V_l^{k-1}\|_F^2,
\end{aligned} \tag{101}$$

$$\begin{aligned}
H_{3,l}^k =\ & (L_1\gamma)^{2(LA-1)}(\beta')_{LA}^2\|V_{LA}^{k-1} - V_{LA}^{k-2}\|_F^2 \\
&+ \sum_{i=l+1}^{LA-1}(L_1\gamma)^{2(i-l)}[(\beta')_i + (L_1^2 + 2L_3C_3\gamma^i \\
&+ L_2C_3\gamma^i)\beta_{i+1}\gamma^2]^2\|V_i^{k-1} - V_i^{k-2}\|_F^2 + (L_1^2 \\
&+ 2L_3C_3\gamma^l + L_2C_3\gamma^l)^2\beta_{l+1}^2\gamma^4\|V_l^{k-1} - V_l^{k-2}\|_F^2,
\end{aligned} \tag{102}$$

$$H_l^k = (L_1\gamma)^{2(LA-l)}\|(U')_{LA}^k - (U')_{LA}^{k-1}\|_F^2. \tag{103}$$

Substituting (92), (93), (94), (98) and (99) into Lemma 2 and after some simplifications yields

$$
\begin{aligned}
&L(Q^k)+\sum_{l=1}^{L}\xi_l\|V_l^k-V_l^{k-1}\|_F^2+\sum_{l=1}^{LA}\delta_l\|d_l^k-d_l^{k-1}\|_F^2\\
&\le L(Q^{k-1})+\sum_{l=1}^{L}\xi_l\|V_l^{k-1}-V_l^{k-2}\|_F^2+\sum_{l=1}^{LA}\delta_l\|d_l^{k-1}-d_l^{k-2}\|_F^2\\
&-\sum_{l=1}^{L}\zeta_l\|W_l^k-W_l^{k-1}\|_F^2-\sum_{l=1}^{L}(\eta_l-\xi_l)\|V_l^k-V_l^{k-1}\|_F^2\\
&-\sum_{l=1}^{LA}(\mu_l-\delta_l)\|d_l^k-d_l^{k-1}\|_F^2. \qquad (104)
\end{aligned}
$$

For the full connection layers, namely for $l=LA+1,...,L$,

$$
\begin{aligned}
\zeta_L&=\frac{\lambda}{2}-3C_3^2\beta_{L-1}^{-1}\beta_L^2\gamma^{2(L-1)}\\
&-2L_1^{-2}C_3^2\beta_L^2\gamma^{2(L-2)}\sum_{l=LA+1}^{L-2}\beta_l^{-1}(L-l)(L_1\gamma)^{2(L-l)}\\
&-2(L-LA)(L_1\gamma)^{2(L-LA)}L_1^{-2}C_3^2\beta_L^2\gamma^{2(L-2)}\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (105)\\
\zeta_i&=\frac{\lambda}{2}-2(1\\
&+3L_1^{-1}L_2L_3\gamma^{i-1})^2C_3^2\beta_i^2\gamma^{2(i-2)}\sum_{l=LA+1}^{i-1}\beta_l^{-1}(L-l)(L_1\gamma)^{2(i-l)}\\
&-2(L-LA)(L_1\gamma)^{2(i-LA)}(1\\
&+3L_1^{-1}L_2L_3\gamma^{i-1})^2C_3^2\beta_i^2\gamma^{2(i-2)}\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (106)\\
\eta_L&=\frac{1+\beta_L}{2}-\beta_L^{-1}-3\gamma^2(1+\beta_L)^2\beta_{L-1}^{-1}\\
&-\frac{4(1+\beta_L)^2}{L_1^2}\sum_{l=LA+1}^{L-2}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l)}\\
&-4(L-LA+1)(L_1\gamma)^{2(L-LA)}(1\\
&+\beta_L^2)^2L_1^{-2}\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (107)\\
\xi_L&=3\gamma^2\beta_L^2\beta_{L-1}^{-1}+\frac{4\beta_L^2}{L_1^2}\sum_{l=LA+1}^{L-2}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l)}\\
&-4(L-LA+1)(L_1\gamma)^{2(L-LA)}\beta_L^2L_1^{-2}\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (108)
\end{aligned}
$$

$$
\begin{aligned}
\eta_{L-1}&=\frac{\beta_{L-1}}{2}-4\beta_{L-1}^2\sum_{l=LA+1}^{L-2}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l-1)}\\
&-4(L-LA+1)(L_1\gamma)^{2(L-LA-1)}\beta_{L-1}^2\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (109)\\
\xi_{L-1}&=4\beta_{L-1}^2\sum_{l=LA+1}^{L-2}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l-1)}\\
&+4(L-LA+1)(L_1\gamma)^{2(L-LA-1)}\beta_{L-1}^2\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (110)
\end{aligned}
$$

and for $l=LA+1,...,L-2$,

$$
\begin{aligned}
\eta_i&=\frac{\beta_i}{2}-4[\beta_i+(L_1^2\\
&+2L_3C_3\gamma^i)\gamma^2\beta_{i+1}]^2\sum_{l=LA+1}^{i-1}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(i-l)}\\
&-4(L_1^2+2L_3C_3\gamma^i)^2\gamma^4\beta_{i+1}^2\beta_i^{-1}(L-i+1)\\
&-4(L-LA+1)(L_1\gamma)^{2(i-LA)}[\beta_i^2+(L_1^2\\
&+2L_3C_3\gamma^i)\gamma^2\beta_{i+1}]\sum_{j=1}^{LA}[4(\beta')_j^{-1}+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (111)\\
\xi_i&=4[\beta_i+(L_1^2+2L_3C_3\gamma^i\\
&+L_2C_3\gamma^i)\gamma^2\beta_{i+1}]^2\sum_{l=LA+1}^{i-1}\beta_l^{-1}(L-l+1)(L_1\gamma)^{2(i-l)}\\
&+4(L_1^2+2L_3C_3\gamma^i+L_2C_3\gamma^i)^2\gamma^4\beta_{i+1}^2\beta_i^{-1}(L-i+1)\\
&+4(L-LA+1)(L_1\gamma)^{2(i-LA)}[\beta_i^2+(L_1^2+2L_3C_3\gamma^i\\
&+L_2C_3\gamma^i)\gamma^2\beta_{i+1}]\sum_{j=1}^{LA}[4(\beta')_j^{-1}+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (112)
\end{aligned}
$$

$$
\begin{aligned}
\eta_{LA+1}&=\frac{\beta_{LA+1}}{2}-4(L_1^2\\
&+2L_3C_3\gamma^{LA+1})^2\gamma^4\beta_{LA+2}^2\beta_{LA+1}^{-1}\\
&-4(L-LA+1)(L_1\gamma)^2[(L_1^2\\
&+2L_3C_3\gamma^{LA})\gamma^2\beta_{LA+1}]^2\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}, \qquad (113)\\
\xi_{LA+1}&=4(L_1^2+2L_3C_3\gamma^{LA+1}\\
&+L_2C_3\gamma^{LA+1})^2\gamma^4\beta_{LA+2}^2\beta_{LA+1}^{-1}+4(L\\
&-LA+1)(L_1\gamma)^2[\beta_i+(L_1^2+2L_3C_3\gamma^{LA}\\
&+L_2C_3\gamma^{LA})\gamma^2\beta_{LA+1}]^2\sum_{j=1}^{LA}[4(\beta')_j^{-1}\\
&+12\beta_j^{-1}](L_1\gamma)^{2(LA-j)}. \qquad (114)
\end{aligned}
$$

For the convolutional layers, namely for $l=1,...,LA$,

$$
\begin{aligned}
\zeta_i&=\frac{\lambda}{2}-4(C_3\beta_i\gamma^{i-2}+3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3})^2\sum_{l=1}^{i-1}[(\beta')_l^{-1}\\
&+3\beta_l^{-1}](LA-l+2)(L_1\gamma)^{2(i-l)}, \qquad (115)\\
\zeta_1&=\frac{\lambda}{2}, \qquad (116)
\end{aligned}
$$

$$\eta_{LA} = \frac{(\beta')_{LA}}{2} - 3\beta_{LA}^{-1}(\beta')_{LA}^2 - 4(\beta')_{LA}^2 \sum_{l=1}^{LA}[(\beta')_l^{-1} + 3\beta_l^{-1}](LA-l+1)(L_1\gamma)^{2(LA-l)}, \tag{117}$$

$$\xi_{LA} = 3\beta_{LA}^{-1}(\beta')_{LA}^2 + 4(\beta')_{LA}^2 \sum_{l=1}^{LA}[(\beta')_l^{-1} + 3\beta_l^{-1}](LA-l+1)(L_1\gamma)^{2(LA-l)}, \tag{118}$$

$$\begin{aligned}\eta_i = {} & \frac{(\beta')_i}{2} - 3\beta_i^{-1}(\beta')_i^2 \\ & - 4[(\beta')_i + (L_1^2 + 2L_3C_3\gamma^i)\gamma^2\beta_{i+1}]^2 \sum_{l=1}^{i-1}[(\beta')_l^{-1} + 3\beta_l^{-1}](LA-l+1)(L_1\gamma)^{2(i-l)} \\ & - 4(L_1^2 + 2L_3C_3\gamma^i)\gamma^4\beta_{i+1}]^2[(\beta')_i^{-1} + 3\beta_i^{-1}](LA-i+1),\end{aligned} \tag{119}$$

$$\begin{aligned}\xi_i = {} & 3\beta_i^{-1}(\beta')_i^2 + 4[(\beta')_i + (L_1^2 + 2L_3C_3\gamma^i \\ & + L_2C_3\gamma^i)\gamma^2\beta_{i+1}]^2 \sum_{l=1}^{i-1}[(\beta')_l^{-1} + 3\beta_l^{-1}](LA-l \\ & + 1)(L_1\gamma)^{2(i-l)} - 4(L_1^2 + 2L_3C_3\gamma^i \\ & + L_2C_3\gamma^i)\gamma^4\beta_{i+1}]^2[(\beta')_i^{-1} \\ & + 3\beta_i^{-1}](LA-i+1),\end{aligned} \tag{120}$$

$$\eta_1 = \frac{(\beta')_1}{2} - 3\beta_1^{-1}(\beta')_1^2 - 4LA(L_1^2 + 2L_3C_3\gamma)\gamma^4\beta_2^2[(\beta')_1^{-1} + 3\beta_1^{-1}], \tag{121}$$

$$\xi_1 = 3\beta_1^{-1}(\beta')_1^2 + 4LA(L_1^2 + 2L_3C_3\gamma + L_2C_3\gamma)\gamma^4\beta_2^2[(\beta')_1^{-1} + 3\beta_1^{-1}], \tag{122}$$

$$\mu_i = \frac{(\beta')_i + \beta_i}{2} - 4\beta_i^2 \sum^{i-1}((\beta')_i^{-1} + 3\beta_i^{-1})(LA-l+2)(L_1\gamma)^{2(i-l)}, \tag{123}$$

$$\delta_i = -4\beta_i^2 \sum^{i-1}((\beta')_i^{-1} + 3\beta_i^{-1})(LA-l+2)(L_1\gamma)^{2(i-l)}. \tag{124}$$

Based on (104), to prove sufficient descent lemma, we need to show that

$$\zeta_i > 0,\ \eta_i - \xi_i > 0,\ i = 1, ..., L \tag{125}$$

$$\mu_i - \delta_i > 0,\ i = 1, ..., LA \tag{126}$$

Then let

$$a = \min\{\zeta_i,\ \eta_i - \xi_i,\ \mu_i - \delta_i\}, \tag{127}$$

we get a in sufficient descent lemma. In the following, we prove (125). Specially, there holds by assumption and the definition of C_3,

$$\frac{\beta_j}{\beta_i} \geq \frac{\beta_j}{(\beta')_i} \geq f_{\min}^{2(i-j)}\gamma^{2(i-j)},\ j < i \leq LA \tag{128}$$

$$\frac{\beta_j}{\beta_i} \geq f_{\min}^{2(i-j)}\gamma^{2(i-j)},\ LA+1 \leq j < i \leq L \tag{129}$$

$$\alpha_3 \geq \max\{24L+1,\ 96LA+1\} \tag{130}$$

It is obvious that, for convolution layers, $\zeta_1 = \frac{\lambda}{2} \geq 0$. By above inequality, for $i = 2, ..., LA$,

$$\begin{aligned}\zeta_i = {} & \frac{\lambda}{2} - 12(C_3\beta_i\gamma^{i-2} + 3L_1^{-1}L_2C_3^2\beta_i\gamma^{2i-3}) \sum_{l=1}^{i-1}[(\beta')_l^{-1} + 3\beta_l^{-1}](LA-l)(L_1\gamma)^{2(i-l)} \\ \geq {} & \frac{\lambda}{2} - 12\beta_i(C_3\gamma^{i-2} + 3L_1^{-1}L_2C_3^2\gamma^{2i-3})^2 \sum_{l=1}^{i-1} 4(LA-l)\alpha_3^{-(i-l)} \\ \geq {} & \frac{\lambda}{2} - 12\beta_i(C_3\gamma^{i-2} + 3L_1^{-1}L_2C_3^2\gamma^{2i-3})^2 \frac{4LA}{\alpha_3 - 1} \\ \geq {} & 0,\end{aligned} \tag{131}$$

where the final inequality is due to (23) and (130). Then we prove $\eta_i - \xi_i > 0$. It is obviously that

$$\begin{aligned}\eta_1 - \xi_1 = {} & \frac{(\beta')_1}{2} - 6\beta_1^{-1}(\beta')_1^2 - 4LA((L_1^2 + 2L_3C_3\gamma)^2 \\ & + (L_1^2 + 2L_3C_3\gamma + L_2C_3\gamma)^2)\gamma^4\beta_2^2[(\beta')_1^{-1} + 3\beta_1^{-1}] \\ \geq {} & \frac{(\beta')_1}{2} - 6\beta_1^{-1}(\beta')_1^2 - 16LA[(L_1^2 + 2L_3C_3\gamma)^2 \\ & + (L_1^2 + 2L_3C_3\gamma + L_2C_3\gamma)^2]\gamma^4\beta_2^2(\beta')_1^{-1} \\ \geq {} & 0,\end{aligned} \tag{132}$$

where the final inequality follows from (20).

For $i = 2, ..., LA-1$, let $\alpha_1 = (L_1^2 + 2L_3C_3\gamma^i)\gamma^2$, $\alpha_2 = (L_1^2 + 2L_3C_3\gamma^i + L_3C_3\gamma^i)\gamma^2$, then there holds,

$$\begin{aligned}& \eta_i - \xi_i \\ & = \frac{(\beta')_i}{2} - 6\beta_i^{-1}(\beta')_i^2 - 4[((\beta')_i + \alpha_1\beta_{i+1})^2 \\ & + ((\beta')_i + \alpha_2\beta_{i+1})^2] \sum_{l=1}^{i-1}((\beta')_1^{-1} + 3\beta_1^{-1})(LA-l \\ & + 1)(L_1\gamma)^{2(i-l)} - 4(\alpha_1^2 + \alpha_2^2)\beta_{i+1}^2((\beta')_i^{-1} \\ & + 3\beta_i^{-1})(LA-i+1) \\ & \geq \frac{(\beta')_i}{2} - 6(\beta')_i^{-1}\beta_{i+1}^2 - 4[((\beta')_i + \alpha_1\beta_{i+1})^2 \\ & + ((\beta')_i + \alpha_2\beta_{i+1})^2] \sum_{l=1}^{i-1} 4(\beta')_1^{-1}(LA-l \\ & + 1)\alpha^{-(i-l)} - 16LA(\alpha_1^2 + \alpha_2^2)\beta_{i+1}^2(\beta')_i^{-1} \\ & \geq (\beta')_i[\frac{1}{2} - \frac{16LA}{\alpha_3 - 1}((1 + \alpha_1\frac{\beta_{i+1}}{(\beta')_i})^2 + (1 + \alpha_2\frac{\beta_{i+1}}{(\beta')_i})^2) \\ & - [16LA(\alpha_1^2 + \alpha_2^2) + 6]\frac{\beta_{i+1}}{(\beta')_i})^2] \\ & \geq \frac{16LA}{\alpha_3 - 1}(\beta')_i[(\frac{1}{2} - 6\frac{(\beta')_i}{\beta_i})\frac{\alpha_3 - 1}{16LA} - 2 \\ & - 2(\alpha_1 + \alpha_2)\frac{\beta_{i+1}}{(\beta')_i} - 2\alpha_3(\alpha_1^2 + \alpha_2^2)(\frac{\beta_{i+1}}{(\beta')_i})^2 \\ & \geq \frac{16LA}{\alpha_3 - 1}(\beta')_i[(\frac{1}{2} - 6\frac{(\beta')_i}{\beta_i})\frac{\alpha_3 - 1}{16LA} - 2 \\ & - 2(\alpha_1 + \alpha_2)\frac{\beta_{i+1}}{(\beta')_i} - 2\alpha_3(\alpha_1 + \alpha_2)^2(\frac{\beta_{i+1}}{(\beta')_i})^2 \\ & \geq 0,\end{aligned} \tag{133}$$

where the inequality follows from (20), (22), (130) and

$$\begin{aligned}&\frac{1+\sqrt{1+2\alpha_3(\frac{\alpha_3-1}{8L}-2)}}{4\alpha_3}\\&\leq 1+\sqrt{1+2\alpha_3(\frac{\alpha_3-1}{8L}-2)}\\&\leq 1+\sqrt{1+10(96LA+1)}\\&\leq 34\sqrt{LA}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\eta_{LA}-\xi_{LA}\\&=\frac{(\beta')_{LA}}{2}-6\beta_{LA}^{-1}(\beta')_{LA}^2\\&-8(\beta')_{LA}^2\sum_{l=1}^{LA}((\beta')_l^{-1}+3\beta_l^{-1})(LA-l+1)(L_1\gamma)^{2(LA-l)}\\&\geq\frac{(\beta')_{LA}}{2}-6\beta_{LA}^{-1}(\beta')_{LA}^2\\&-8(\beta')_{LA}\sum_{l=1}^{LA}(LA-l+1)\alpha_3^{-(LA-l)}\\&\geq(\beta')_{LA}(\frac{1}{2}-\frac{32LA}{\alpha_3-1}-6(\frac{(\beta')_{LA}}{\beta_{LA}})^2)\\&\geq 0,\end{aligned}\tag{134}$$

where the inequality follows from (20) and (130).

For the output of convolution layers,

$$\begin{aligned}\mu_i-\delta_i&=\frac{(\beta')_i+\beta_i}{2}\\&-8\beta_i^2\sum_{l=1}^{i-1}((\beta')_l^{-1}+3\beta_1^{-1})(LA-l+2)(L_1\gamma)^{2(i-l)}\\&\geq\frac{(\beta')_i+\beta_i}{2}-8\beta_i\sum_{l=1}^{i-1}4(LA-l+2)\alpha_3^{-(i-l)}\\&\geq\frac{(\beta')_i+\beta_i}{2}-\frac{32LA}{\alpha_3-1}\beta_i\\&\geq\frac{1}{2}+\frac{\beta_i}{(2\beta')_i}-\frac{32LA}{\alpha_3-1}\frac{\beta_i}{(\beta')_i}\\&\geq 0,\end{aligned}\tag{135}$$

where the final inequality follows from $\frac{32LA}{\alpha_3-1}\geq\frac{1}{3}$.

For the full connection layers, namely for $l=LA+1,...,L$,

$$\begin{aligned}&\zeta_i\\&\geq\frac{\lambda}{2}-2C_3^2\beta_i\gamma^{2(i-2)}(1+\\&+3L_1^{-1}L_2L_3\gamma^{i-1})^2\sum_{l=LA+1}^{i-1}(L-l)\alpha^{-(i-l)}\\&-2(L-LA)(L_1\gamma)^{2(i-LA)}(1\\&+3L_1^{-1}L_2L_3\gamma^{i-1})^2C_3^2\beta_i\gamma^{2(i-2)}\sum_{j=1}^{LA}16\frac{\beta_i}{(\beta')_j}(L_1\gamma)^{2(LA-j)}\\&>\frac{\lambda}{2}-2C_3^2\beta_i\gamma^{2(i-2)}(1+\\&+3L_1^{-1}L_2L_3\gamma^{i-1})^2\sum_{l=LA+1}^{i-1}(L-l)\alpha^{-(i-l)}\\&-2(L-LA)(L_1\gamma)^{2(i-LA)}(1\\&+3L_1^{-1}L_2L_3\gamma^{i-1})^2C_3^2\beta_i\gamma^{2(i-2)}\sum_{j=1}^{LA}\alpha_3^{-(LA-j)}\\&>\frac{\lambda}{2}-2C_3^2\beta_i\gamma^{2(i-2)}(1+3L_1^{-1}L_2L_3\gamma^{i-1})^2\frac{L}{\alpha_3-1}\\&-2(L-LA)(L_1\gamma)^{2(i-LA)}(1\\&+3L_1^{-1}L_2L_3\gamma^{i-1})^2C_3^2\beta_i\gamma^{2(i-2)}\frac{1}{\alpha_3-1}\\&>\frac{\lambda}{2}-2C_3^2\beta_i\gamma^{2(i-2)}(1+3L_1^{-1}L_2L_3\gamma^{i-1})^2[\frac{L}{\alpha_3-1}\\&+\frac{(L-LA)(L_1\gamma)^{2(i-LA)}}{\alpha_3-1}]\\&>0,\end{aligned}\tag{136}$$

where the final inequality follows from (23).

Similarly, we have

$$\begin{aligned}&\zeta_L\\&\geq\frac{\lambda}{2}-\beta_LC_3^2\gamma^{2(L-2)}(\frac{3}{16}+\frac{1}{8}\sum_{l=LA+1}^{L-1}(L-l)\alpha^{-(L-l-1)})\\&-32(L-LA)(L_1\gamma)^{2(L-LA)}L_1^{-2}\beta_LC_3^2\gamma^{2(L-2)}\sum_{j=1}^{LA}\alpha_3^{-(LA-j-1)}\\&>\frac{\lambda}{2}-\beta_LC_3^2\gamma^{2(L-2)}(\frac{3}{16}+\frac{1}{8(\alpha_3-1)})\\&-32(L-LA)(L_1\gamma)^{2(L-LA)}L_1^{-2}\beta_LC_3^2\gamma^{2(L-2)}\frac{1}{\alpha_3-1}\\&>\frac{\lambda}{2}-\beta_LC_3^2\gamma^{2(L-2)}(\frac{3}{16}+\frac{1}{8(\alpha_3-1)})\\&-32(L_1\gamma)^{2(L-LA)}L_1^{-2}\beta_LC_3^2\gamma^{2(L-2)}\frac{L}{\alpha_3-1}\\&>\frac{\lambda}{2}-\beta_LC_3^2\gamma^{2(L-2)}(\frac{3}{16}+\frac{1}{8(\alpha_3-1)}\\&+32(L_1\gamma)^{2(L-LA)}L_1^{-2}\frac{1}{\alpha_3-1})\\&>0,\end{aligned}\tag{137}$$

where the final inequality follows from (23).

Similarly, we have,

$$
\begin{aligned}
&\eta_i - \xi_i \\
&\geq \frac{\beta_i}{2} - 4[(\beta_i + \alpha_1\beta_{i+1})^2 \\
&+ (\beta_i + \alpha_2\beta_{i+1})^2] \sum_{l=LA+1}^{i-1} \beta_i^{-1}(L-l+1)\alpha_3^{-(i-l)} \\
&- 4(\alpha_1^2+\alpha_1^2)\beta_{i+1}^2\beta_i^{-1} \\
&- 4(L-LA+1)(L_1\gamma)^{2(i-LA)}[(\beta_i + \alpha_1\beta_{i+1})^2 \\
&+ (\beta_i + \alpha_2\beta_{i+1})^2] \sum_{l=1}^{LA} 16(\beta')_l^{-1}\alpha_3^{-(i-l)} \\
&\geq \beta_i\{\frac{1}{2} - \frac{4L}{\alpha_3-1}[(1+\alpha_1\frac{\beta_{i+1}}{\beta_i})^2 \\
&+ (1+\alpha_2\frac{\beta_{i+1}}{\beta_i})^2] - 4L(\alpha_1^2+\alpha_1^2)(\frac{\beta_{i+1}}{\beta_i})^2 \\
&- \frac{4L}{\alpha_3-1}\phi[(1+\alpha_1\frac{\beta_{i+1}}{\beta_i})^2 + (1+\alpha_2\frac{\beta_{i+1}}{\beta_i})^2]\} \\
&= \frac{4L}{\alpha_3-1}\beta_i[\frac{\alpha_3-1}{8L} - (1+\phi)[2+2(\alpha_1+\alpha_2)\frac{\beta_{i+1}}{\beta_i} \\
&+ (\alpha_1^2+\alpha_1^2)(\frac{\beta_{i+1}}{\beta_i})^2] - (\alpha_3-1)(\alpha_1^2+\alpha_1^2)(\frac{\beta_{i+1}}{\beta_i})^2 \\
&\geq \frac{4L}{\alpha_3-1}\beta_i[\frac{\alpha_3-1}{8L} - 2(1+\phi) - 2(1+\phi)(\alpha_1+\alpha_2)\frac{\beta_{i+1}}{\beta_i} \\
&- (\alpha_3+\phi)(\alpha_1^2+\alpha_1^2)(\frac{\beta_{i+1}}{\beta_i})^2 \\
&\geq 0, \qquad (138)
\end{aligned}
$$

where the final inequality follows from (23) and

$$
\begin{aligned}
&\frac{(1+\phi) + \sqrt{(1+\phi)^2 + (\alpha_3+\phi)(\frac{\alpha_3-1}{8L} - 2(1+\phi))}}{\frac{\alpha_3-1}{8L} - 2(1+\phi)} \\
&\leq (1+\phi) + \sqrt{(1+\phi)^2 + (\alpha_3+\phi)(\frac{\alpha_3-1}{8L} - 2(1+\phi))} \\
&\leq (1+\phi) + \sqrt{(1+\phi)^2 + (1+\phi) + 24L + 16L\phi} \\
&\leq 2(1+\phi) + \sqrt{24L + 16L\phi}.
\end{aligned}
$$

And for the $L-1$ layer, there holds

$$
\begin{aligned}
&\eta_{L-1} - \xi_{L-1} \\
&= \frac{\beta_{L-1}}{2} - 8\beta_{L-1}^2 \sum_{l=LA+1}^{L-2} \beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l+1)} \\
&- 8(L-LA+1)(L_1\gamma)^{2(L-LA-1)}\beta_{L-1}^2 \sum_{j=1}^{LA}[4(\beta')_j^{-1} \\
&+ 12\beta_j^{-1}](L_1\gamma)^{2(LA-j)} \\
&\geq \frac{\beta_{L-1}}{2} - 8\beta_{L-1}^2 \sum_{l=LA+1}^{L-2} \beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l+1)} \\
&- 8(L-LA+1)(L_1\gamma)^{2(L-LA-1)}\beta_{LA}^2 \sum_{j=1}^{LA}[4(\beta')_j^{-1} \\
&+ 12\beta_j^{-1}](L_1\gamma)^{2(LA-j)} \\
&\geq \frac{\beta_{L-1}}{2} - 8\beta_{L-1}^2 \sum_{l=LA+1}^{L-2} (L-l+1)\alpha_3^{-(L-l+1)} - 8(L \\
&- LA+1)(L_1\gamma)^{2(L-LA-1)}\beta_{LA} \sum_{j=1}^{LA} 16\alpha_3^{(LA-j)} \\
&\geq \beta_{L-1}(\frac{1}{2} - \frac{8L}{\alpha_3-1} - \frac{128L}{\alpha_3-1}(L_1\gamma)^{2(L-LA-1)}\frac{\beta_{LA}}{\beta_{L-1}}) \\
&\geq 0, \qquad (139)
\end{aligned}
$$

where the final inequality holds for (21).

$$
\begin{aligned}
&\eta_L - \xi_L \\
&= \frac{1+\beta_L}{2} - \beta_L^{-1} - 3\gamma^2\beta_{L-1}^{-1}[(1+\beta_L)^2 + \beta_L^2] \\
&- \frac{4}{L_1^2}[(1+\beta_L)^2 + \beta_L^2] \sum_{l=LA+1}^{L-2} \beta_l^{-1}(L-l+1)(L_1\gamma)^{2(L-l)} \\
&- 4(L-LA+1)(L_1\gamma)^{2(L-LA)}[(1+\beta_L)^2 \\
&+ \beta_L^2]L_1^{-2} \sum_{j=1}^{LA}[4(\beta')_j^{-1} + 12\beta_j^{-1}](L_1\gamma)^{2(LA-j)} \\
&\geq \frac{1+\beta_L}{2} - \beta_L^{-1} - 3\gamma^2\beta_{L-1}^{-1}[(1+\beta_L)^2 + \beta_L^2] \\
&- 4\beta_{L-1}^{-1}\gamma^2[(1+\beta_L)^2 + \beta_L^2] \sum_{l=LA+1}^{L-2} (L-l+1)\alpha^{-(L-l-1)} \\
&- 4(L-LA+1)(L_1\gamma)^{2(L-LA)}[(1+\beta_L)^2 \\
&+ \beta_L^2]L_1^{-2} \sum_{j=1}^{LA} 16\beta_{LA}^{-1}\alpha_3^{-(LA-j)} \\
&> \frac{\beta_L^2+\beta_L}{2\beta_{L-1}} - 2(\beta_L^2+\beta_L+1)[(3+\frac{4L}{\alpha_3-1})\beta_{L-1}^{-1}\gamma^2 \\
&+ \frac{64}{\alpha_3-1}(L-LA+1)(L_1\gamma)^{2(L-LA)}L_1^{-2}\beta_{LA}^{-1}] \\
&> \frac{\beta_L^2+\beta_L}{2\beta_{L-1}} - 2(\beta_L^2+\beta_L+1)[\frac{19}{6}\beta_{L-1}^{-1}\gamma^2 \\
&+ \frac{8}{3}(L_1\gamma)^{2(L-LA)}L_1^{-2}\beta_{LA}^{-1}] \\
&\geq 0, \qquad (140)
\end{aligned}
$$

where the final inequality holds for (16) and

$$\frac{(\beta')_{LA}}{\beta_L} \leq (L_1\gamma)^{2(L-LA)}\alpha^{L-LA}.$$

This completes the proof. ■

5) Relative Error Lemma: In the following, we will give the lemma which indicate the gradients of the augmented Lagrangian and the new Lyapunov function can be bounded by the discrepancy between two successive updates.

Lemma 8: Under conditions of Assumption 1, for any positive $k \geq 2$, there exists some positive constant $\bar{b}$ such that

$$\begin{aligned}\|\nabla L(Q^k)\|_F \leq \bar{b}((&\sum_{l=1}^{L}\|W_l^k - W_l^{k-1}\|_F + \|V_l^k - V_l^{k-1}\|_F \\ &+ \|V_l^{k-1} - V_l^{k-2}\|_F) + \sum_{l=1}^{LA}(\|d_l^k - d_l^{k-1}\|_F \\ &+ \|d_l^{k-1} - d_l^{k-2}\|_F)),\end{aligned} \tag{141}$$

and $\|\nabla\hat{L}(\hat{Q}^k)\|_F \leq \hat{b}\|\hat{Q}^k - \hat{Q}^{k-1}\|_F$, where $\hat{b} = \sqrt{3L + 2LA}b$, $b = \bar{b} + 4\Omega_{max}$, $\Omega_{max} = \{\xi_l,\ \delta_l\}$.

Proof: Note that

$$\begin{aligned}\|\nabla L(Q^k)\|_F =&(\{\frac{\partial L(Q^k)}{\partial W_l}\}_{l=1}^{L}, \{\frac{\partial L(Q^k)}{\partial V_l}\}_{l=1}^{L}, \\ &\{\frac{\partial L(Q^k)}{\partial d_l}\}_{l=1}^{LA}, \{\frac{\partial L(Q^k)}{\partial U_l}\}_{l=1}^{L}, \{\frac{\partial L(Q^k)}{\partial (U')_l}\}_{l=1}^{LA}),\end{aligned} \tag{142}$$

then

$$\begin{aligned}\|\nabla L(Q^k)\|_F \leq \sum_{l=1}^{L}(&\|\frac{\partial L(Q^k)}{\partial W_l}\|_F + \|\frac{\partial L(Q^k)}{\partial V_l}\|_F \\ &+ \|\frac{\partial L(Q^k)}{\partial d_l}\|_F + \|\frac{\partial L(Q^k)}{\partial U_l}\|_F + \|\frac{\partial L(Q^k)}{\partial (U')_l}\|_F).\end{aligned} \tag{143}$$

We need to bound each component of the $\nabla L(Q^k)$ to bound the gradients of the augmented Lagrangian.

For the weight parameters, by the optimality condition of W_L^k,

$$\begin{aligned}&\lambda W_L^k + \beta_L(W_L^k V_{L-1}^{k-1} - V_L^{k-1})(V_L^{k-1})^T + U_L^{k-1}(V_L^{k-1})^T \\ &= 0,\end{aligned} \tag{144}$$

which implies,

$$\begin{aligned}\frac{\partial L(Q^k)}{\partial W_L} &= \lambda W_L^k + \beta_L(W_L^k V_{L-1}^k - V_L^k)(V_L^k)^T + U_L^k(V_L^k)^T \\ &= \beta_L[(W_L^k V_{L-1}^k - V_L^k)(V_{L-1}^k - V_{L-1}^{k-1})^T \\ &+ (W_L^k(V_{L-1}^k - V_{L-1}^{k-1}) - (V_L^k - V_L^{k-1}))V_{L-1}^{k-1}]^T \\ &+ U_L^{k-1}(V_{L-1}^k - V_L^{k-1})^T \\ &+ (U_L^k - U_L^{k-1})(V_{L-1}^k)^T.\end{aligned} \tag{145}$$

By the boundedness of 6, the above equality yields

$$\begin{aligned}\|\frac{\partial L(Q^k)}{\partial W_L}\| \leq &10\beta_L C_3\gamma^{N-1}\|V_{L-1}^k - V_{L-1}^{k-1}\|_F \\ &+ 3C_3\gamma^{N-2}(\beta_L + 1)\|V_L^k - V_L^{k-1}\|_F.\end{aligned} \tag{146}$$

By the optimality condition of W_i^k,

$$\begin{aligned}&\lambda W_i^k + ((\beta_i\sigma((W_i^{k-1}V_{i-1}^{k-1}) - \beta_i V_i^{k-1} \\ &+ U_i^{k-1}) \odot \sigma'(W_i^{k-1}V_{i-1}^{k-1})))(V_{i-1}^{k-1})^T \\ &+ \frac{\beta_i h_i^k}{2}(W_i^k - W_i^{k-1})(V_{i-1}^{k-1})^T = 0,\end{aligned} \tag{147}$$

which implies

$$\begin{aligned}&\frac{\partial L(Q^k)}{\partial W_l} \\ &= \lambda W_i^k + ((\beta_i\sigma(W_i^k V_{i-1}^k) - \beta_i V_i^k \\ &+ U_i^k) \odot \sigma'(W_i^k V_{i-1}^k))(V_{i-1}^k)^T \\ &= ((\beta_i\sigma((W_i^k V_{i-1}^k) - \beta_i V_i^k + U_i^k) \odot \sigma'(W_i^k V_{i-1}^k)))(V_{i-1}^k)^T \\ &- ((\beta_i\sigma((W_i^{k-1}V_{i-1}^{k-1}) - \beta_i V_i^{k-1} \\ &+ U_i^{k-1}) \odot \sigma'(W_i^{k-1}V_{i-1}^{k-1})))(V_{i-1}^{k-1})^T \\ &- \frac{\beta_i h_i^k}{2}(W_i^k - W_i^{k-1})(V_{i-1}^{k-1})^T \\ &= [\beta_i(\sigma(W_i^k V_{i-1}^k) - \sigma(W_i^{k-1}V_{i-1}^k) \\ &+ \sigma(W_i^{k-1}V_{i-1}^k) - \sigma(W_i^{k-1}V_{i-1}^{k-1})) \odot \sigma'(W_i^k V_{i-1}^k) \\ &+ (\beta_i(V_i^{k-1} - V_i^k) + (U_i^k - U_i^{k-1}))\sigma'(W_i^k V_{i-1}^k) \\ &+ (\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1}) - \beta_i V_i^{k-1} \\ &+ U_i^{k-1}) \odot (\sigma'(W_i^k V_{i-1}^k) - \sigma'(W_i^{k-1}V_{i-1}^k)) \\ &+ (\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1}) - \beta_i V_i^{k-1} \\ &+ U_i^{k-1}) \odot (\sigma'(W_i^{k-1}V_{i-1}^k) - \sigma'(W_i^{k-1}V_{i-1}^{k-1}))](V_{i-1}^k)^T \\ &+ (\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1}) - \beta_i V_i^{k-1} \\ &+ U_i^{k-1}) \odot \sigma'(W_i^{k-1}V_{i-1}^{k-1})(V_{i-1}^k - V_{i-1}^{k-1})^T \\ &- \frac{\beta_i h_i^k}{2}(W_i^k - W_i^{k-1})(V_{i-1}^{k-1})^T.\end{aligned} \tag{148}$$

By Assumption 1 and Lemma 6, the above equality yields

$$\begin{aligned}&\|\frac{\partial L(Q^k)}{\partial W_l}\| \\ &\leq 3C_3\gamma^{i-2}[3\beta_i C_3\gamma^{i-2}(L_1^2 + L_0L_2\sqrt{nm_i} \\ &+ 4L_2C_3\gamma^{i-1} + \frac{2}{3}L_3\gamma)\|W_i^k - W_i^{k-1}\|_F] \\ &+ \beta_i[L_1^2\gamma + (L_1 + L_2\gamma)(L_0\sqrt{nm_i} \\ &+ 4L_2C_3\gamma^{i-1})]\|V_{i-1}^k - V_{i-1}^{k-1}\|_F \\ &+ \beta_i L_1\|V_i^k - V_i^{k-1}\|_F + L_1\|U_i^k - U_i^{k-1}\|_F].\end{aligned} \tag{149}$$

Similarly, for the weight parameters in convolution layers, by the optimality condition of W_i^k,

$$\begin{aligned}&\lambda W_i^k + ((\beta_i\sigma((W_i^{k-1}V_{i-1}^{k-1}) - \beta_i d_i^{k-1} \\ &+ U_i^{k-1}) \odot \sigma'(W_i^{k-1}V_{i-1}^{k-1})))(V_{i-1}^{k-1})^T \\ &+ \frac{\beta_i h_i^k}{2}(W_i^k - W_i^{k-1})(V_{i-1}^{k-1})^T = 0,\end{aligned} \tag{150}$$

which implies

$$\begin{aligned}
&\frac{\partial L(Q^k)}{\partial W_l}\\
&=\lambda W_i^k+((\beta_i\sigma(W_i^kV_{i-1}^k)-\beta_i d_i^k\\
&\quad+U_i^k)\odot\sigma'(W_i^kV_{i-1}^k))(V_{i-1}^k)^T\\
&=((\beta_i\sigma((W_i^kV_{i-1}^k)-\beta_i d_i^k\\
&\quad+U_i^k)\odot\sigma'(W_i^kV_{i-1}^k)))(V_{i-1}^k)^T\\
&\quad-((\beta_i\sigma((W_i^{k-1}V_{i-1}^{k-1})-\beta_i d_i^{k-1}\\
&\quad+U_i^{k-1})\odot\sigma'(W_i^{k-1}V_{i-1}^{k-1})))(V_{i-1}^{k-1})^T\\
&\quad-\frac{\beta_i h_i^k}{2}(W_i^k-W_i^{k-1})(V_{i-1}^{k-1})^T\\
&=[\beta_i(\sigma(W_i^kV_{i-1}^k)-\sigma(W_i^{k-1}V_{i-1}^k)\\
&\quad+\sigma(W_i^{k-1}V_{i-1}^k)-\sigma(W_i^{k-1}V_{i-1}^{k-1}))\odot\sigma'(W_i^kV_{i-1}^k)\\
&\quad+(\beta_i(d_i^{k-1}-d_i^k)+(U_i^k-U_i^{k-1}))\sigma'(W_i^kV_{i-1}^k)\\
&\quad+(\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1})-\beta_i d_i^{k-1}\\
&\quad+U_i^{k-1})\odot(\sigma'(W_i^kV_{i-1}^k)-\sigma'(W_i^{k-1}V_{i-1}^k))\\
&\quad+(\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1})-\beta_i d_i^{k-1}\\
&\quad+U_i^{k-1})\odot(\sigma'(W_i^{k-1}V_{i-1}^k)-\sigma'(W_i^{k-1}V_{i-1}^{k-1}))](V_{i-1}^k)^T\\
&\quad+(\beta_i\sigma(W_i^{k-1}V_{i-1}^{k-1})-\beta_i d_i^{k-1}\\
&\quad+U_i^{k-1})\odot\sigma'(W_i^{k-1}V_{i-1}^{k-1})(V_{i-1}^k-V_{i-1}^{k-1})^T\\
&\quad-\frac{\beta_i h_i^k}{2}(W_i^k-W_i^{k-1})(V_{i-1}^{k-1})^T. \qquad (151)
\end{aligned}$$

By Assumption 1 and Lemma 6, the above equality yields

$$\begin{aligned}
&\|\frac{\partial L(Q^k)}{\partial W_l}\|\\
&\le 3C_3\gamma^{i-2}[3\beta_i C_3\gamma^{i-2}(L_1^2+L_0L_2\sqrt{nm_i}\\
&\quad+4L_2C_3\gamma^{i-1}+\frac{2}{3}L_3\gamma)\|W_i^k-W_i^{k-1}\|_F]\\
&\quad+\beta_i[L_1^2\gamma+(L_1+L_2\gamma)(L_0\sqrt{nm_i}\\
&\quad+4L_2C_3\gamma^{i-1})]\|V_{i-1}^k-V_{i-1}^{k-1}\|_F\\
&\quad+\beta_iL_1\|d_i^k-d_i^{k-1}\|_F+L_1\|U_i^k-U_i^{k-1}\|_F]. \qquad (152)
\end{aligned}$$

For the weight parameters in input layer, by the optimality condition of W_1^k,

$$\begin{aligned}
&\lambda W_1^k+((\beta_1\sigma((W_1^{k-1}X)-\beta_1 d_1^{k-1}\\
&+U_1^{k-1})\odot\sigma'(W_1^{k-1}X)))X^T\\
&+\frac{\beta_1 h_1^k}{2}(W_1^k-W_1^{k-1})X^T=0, \qquad (153)
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{\partial L(Q^k)}{\partial W_1}&=\lambda W_1^k+[(\beta_1\sigma(W_1^kX)-\beta_1 d_1^k\\
&\quad+U_1^k)\odot\sigma'(W_1^kX)]X^T\\
&=[(\beta_1(\sigma(W_1^kX)-\sigma(W_1^{k-1}X))-\beta_1(d_1^k-d_1^{k-1})\\
&\quad+(U_1^k-U_1^{k-1}))\odot\sigma(W_1^kX)]X^T\\
&\quad+[(\beta_1\sigma(W_1^{k-1}X)-\beta_1 d_1^{k-1}\\
&\quad+U_1^{k-1})\odot(\sigma'(W_1^kX)-\sigma'(W_1^{k-1}X))]X^T\\
&\quad+\frac{\beta_1 h_1^k}{2}(W_1^{k-1}-W_1^k)X^T. \qquad (154)
\end{aligned}$$

Then there holds,

$$\begin{aligned}
\|\frac{\partial L(Q^k)}{\partial W_1}\|&\le\beta_1\|X\|_F(\|X\|_F(L_1^2+L_0L_2\sqrt{nm_1}\\
&\quad+4L_2C_3)+2L_3C_3)\|W_1^k-W_1^{k-1}\|_F\\
&\quad+\beta_1L_1\|X\|_F\|d_1^k-d_1^{k-1}\|_F\\
&\quad+L_1\|X\|_F\|U_1^k-U_1^{k-1}\|_F. \qquad (155)
\end{aligned}$$

For the output of convolution layers, by the optimality condition of d_1^k, there holds,

$$\begin{aligned}
&\beta_l(\sigma(W_l^kV_{l-1}^k)-d_l^{k-1}+\frac{U_l^{k-1}}{\beta_l})\\
&+(\beta')_l(pool(d_l^{k-1})-V_l^{k-1}+\frac{(U')_l^{k-1}}{(\beta')_l})=0, \qquad (156)
\end{aligned}$$

which implies,

$$\begin{aligned}
\frac{\partial L(Q^k)}{\partial d_1}&=\beta_1(\sigma(W_l^kV_{l-1}^k)-d_l^k+\frac{U_l^k}{\beta_l})\\
&\quad+(\beta')_l(pool(d_l^k)-V_l^k+\frac{(U')_l^k}{(\beta')_l})\\
&=\beta_1(d_l^{k-1}-d_l^k)+(U_l^k-U_l^{k-1})\\
&\quad+(\beta')_l(pool(d_l^k)-pool(d_l^{k-1}))\\
&\quad+(\beta')_l(V_l^{k-1}-V_l^k)+((U')_l^k-(U')_l^{k-1}). \qquad (157)
\end{aligned}$$

Then there holds,

$$\begin{aligned}
&\|\frac{\partial L(Q^k)}{\partial d_1}\|_F\\
&\le\beta_1\|d_l^{k-1}-d_l^k\|_F+\|U_l^k-U_l^{k-1}\|_F\\
&\quad+(\beta')_l\|d_l^k-d_l^{k-1}\|_F+(\beta')_l\|V_l^k-V_l^{k-1}\|_F\\
&\quad+\|(U')_l^k-(U')_l^{k-1}\|_F\\
&=((\beta')_l-\beta_1)\|d_l^k-d_l^{k-1}\|_F-(\beta')_l\|V_l^k-V_l^{k-1}\|_F\\
&\quad+\|U_l^k-U_l^{k-1}\|_F+\|(U')_l^k-(U')_l^{k-1}\|_F. \qquad (158)
\end{aligned}$$

For the output of pooling layers, by the optimality condition of V_1^k, there holds,

$$\begin{aligned}
&\beta_l(V_l^k-pool(d_l^k))+(W_{l+1}^k)^T[(U_{l+1}^{k-1}+\beta_{l+1}(\sigma(W_{l+1}^kV_l^{k-1})\\
&-d_{l+1}^{k-1}))\odot\sigma'(W_{l+1}^kV_l^{k-1})]-(U')_l^{k-1}\\
&+\frac{\beta_{l+1}t_1^k}{2}W_{l+1}^k(W_{l+1}^k)^T(V_1^k-V_1^{k-1})=0, \qquad (159)
\end{aligned}$$

which implies,

$$\begin{aligned}
&\frac{\partial L(Q^k)}{\partial V_1}\\
&=(W_{l+1}^k)^T[((U_{l+1}^k-U_{l+1}^{k-1})\\
&+\beta_{l+1}(\sigma(W_{l+1}^kV_l^k)-\sigma(W_{l+1}^kV_l^{k-1}))\\
&+\beta_{l+1}(d_{l+1}^{k-1}-d_{l+1}^k))\sigma'(W_{l+1}^kV_l^k)]\\
&+(W_{l+1}^k)^T[(U_{l+1}^{k-1}+\beta_{l+1}(\sigma(W_{l+1}^kV_l^{k-1})\\
&-d_{l+1}^{k-1}))\odot(\sigma'(W_{l+1}^kV_l^k)-\sigma'(W_{l+1}^kV_l^{k-1}))]\\
&+((U')_l^{k-1}-(U')_l^k)+\frac{\beta_{l+1}t_1^k}{2}W_{l+1}^k(W_{l+1}^k)^T(V_1^{k-1}-V_1^k). \qquad (160)
\end{aligned}$$

Then there holds,

$$\begin{aligned}&\|\frac{\partial L(Q^k)}{\partial V_1}\|_F\\&\leq \beta_{l+1}\gamma^2(2C_3\gamma^l(2L_2+L_3)\\&+L_0L_2\sqrt{nm_{l+1}})\|V_1^k-V_1^{k-1}\|_F\\&+\beta_{l+1}L_1\gamma\|d_{l+1}^k-d_{l+1}^{k-1}\|_F\\&+\|(U')_l^k-(U')_l^{k-1}\|_F+L_1\gamma\|U_{l+1}^k-U_{l+1}^{k-1}\|_F,\end{aligned}\tag{161}$$

For the output of fully connected layers, by the optimality condition of V_1^k, there holds,

$$\begin{aligned}&\beta_l(V_l^k-\sigma(W_l^kV_{l-1}^k))+(W_{l+1}^k)^T[(U_{l+1}^{k-1}\\&+\beta_{l+1}(\sigma(W_{l+1}^kV_l^{k-1})-V_{l+1}^{k-1}))\odot\sigma'(W_{l+1}^kV_l^{k-1})]\\&-U_l^{k-1}+\frac{\beta_{l+1}t_1^k}{2}W_{l+1}^k(W_{l+1}^k)^T(V_1^k-V_1^{k-1})=0,\end{aligned}\tag{162}$$

which implies,

$$\begin{aligned}&\frac{\partial L(Q^k)}{\partial V_l}\\&=(W_{l+1}^k)^T[((U_{l+1}^k-U_{l+1}^{k-1})\\&+\beta_{l+1}(\sigma(W_{l+1}^kV_l^k)-\sigma(W_{l+1}^kV_l^{k-1}))\\&+\beta_{l+1}(V_{l+1}^{k-1}-V_{l+1}^k))\sigma'(W_{l+1}^kV_l^k)]\\&+(W_{l+1}^k)^T[(U_{l+1}^{k-1}+\beta_{l+1}(\sigma(W_{l+1}^kV_l^{k-1})\\&-V_{l+1}^{k-1}))\odot(\sigma'(W_{l+1}^kV_l^k)-\sigma'(W_{l+1}^kV_l^{k-1}))]\\&+(U_l^{k-1}-U_l^k)+\frac{\beta_{l+1}t_1^k}{2}W_{l+1}^k(W_{l+1}^k)^T(V_1^{k-1}-V_1^k).\end{aligned}\tag{163}$$

Then there holds,

$$\begin{aligned}&\|\frac{\partial L(Q^k)}{\partial V_1}\|_F\\&\leq\beta_{l+1}\gamma^2(2C_3\gamma^l(2L_2+L_3)\\&+L_0L_2\sqrt{nm_{l+1}})\|V_1^k-V_1^{k-1}\|_F\\&+\beta_{l+1}L_1\gamma\|V_{l+1}^k-V_{l+1}^{k-1}\|_F+\|U_l^k-U_l^{k-1}\|_F\\&+L_1\gamma\|U_{l+1}^k-U_{l+1}^{k-1}\|_F.\end{aligned}\tag{164}$$

On $\|\frac{\partial L(Q^k)}{\partial V_{L-1}}\|_F$, by the optimality condition of V_{L-1}^k,

$$\begin{aligned}&\beta_{L-1}(V_L^k-\sigma(W_{L-1}^kV_{L-2}^k))-U_{L-1}^{k-1}\\&+(W_L^k)^T(U_L^{k-1}+\beta_L(W_L^kV_{L-1}^k-V_L^{k-1}))=0,\end{aligned}\tag{165}$$

which implies

$$\begin{aligned}\frac{\partial L(Q^k)}{\partial V_{L-1}}&=\beta_{L-1}(V_L^k-\sigma(W_{L-1}^kV_{L-2}^k))-U_{L-1}^k\\&\quad+(W_L^k)^T(U_L^k+\beta_L(W_L^kV_{L-1}^k-V_L^k))\\&=U_{L-1}^{k-1}-U_{L-1}^k+(W_L^k)^T(U_L^k-U_L^{k-1})\\&\quad+\beta_L(W_L^k)^T(V_L^{k-1}-U_L^k).\end{aligned}\tag{166}$$

The above equality implies

$$\begin{aligned}\|\frac{\partial L(Q^k)}{\partial V_{L-1}}\|_F&\leq\beta_L\gamma\|V_L^k-V_L^{k-1}\|_F+\|U_{L-1}^k-U_{L-1}^{k-1}\|_F\\&\quad+\gamma\|U_L^k-U_L^{k-1}\|_F.\end{aligned}\tag{167}$$

On $\|\frac{\partial L(Q^k)}{\partial V_L}\|_F$, similarly, by the optimality condition of V_L^k,

$$\|\frac{\partial L(Q^k)}{\partial V_L}\|=\|U_L^k-U_L^{k-1}\|_F.\tag{168}$$

By the update of U_l^k, we can easily get,

$$\|\frac{\partial L(Q^k)}{\partial U_l}\|=\beta_l^{-1}\|U_l^k-U_l^{k-1}\|_F.\tag{169}$$

Substituting the above inequality into (143), and after some simplifications, we get

$$\begin{aligned}\|L(Q^k)\|_F&\leq\overline{\alpha}(\sum_{l=1}^{L}(\|W_l^k-W_l^{k-1}\|_F+\|V_l^k-V_l^{k-1}\|_F\\&+\|U_l^k-U_l^{k-1}\|_F)+\sum_{l=1}^{LA}(\|d_l^k-d_l^{k-1}\|_F\\&+\|(U')_l^k-(U')_l^{k-1}\|_F)),\end{aligned}\tag{170}$$

for some $\overline{\alpha}>0$. By Lemma 7, substituting these upper bounds of $\|U_i^k-U_i^{k-1}\|_F$ into (170) and after some simplifications implies (141) for some constant $\overline{b}$.

By (141), it is easy to derive,

$$\begin{aligned}&\|\nabla\hat{L}(\hat{Q}^k)\|_F\\&\leq\|\nabla L(Q^k)\|_F+\sum_{l=1}^{L}4\xi_l\|V_l^k-V_l^{k-1}\|_F\\&+\sum_{l=1}^{LA}4\delta_l\|d_l^k-d_l^{k-1}\|_F\\&\leq b(\sum_{l=1}^{L}(\|W_l^k-W_l^{k-1}\|_F+\|V_l^k-V_l^{k-1}\|_F\\&+\|V_l^{k-1}-V_l^{k-2}\|_F)+\sum_{l=1}^{LA}(\|d_l^k-d_l^{k-1}\|_F\\&+\|d_l^{k-1}-d_l^{k-2}\|_F))\\&\leq\hat{b}\|\hat{Q}_l^k-\hat{Q}_l^{k-1}\|_F,\end{aligned}\tag{171}$$

where $b=\overline{b}+4\Omega_{\max}$, $\Omega_{\max}=\max\{\xi_l,\ \delta_l\}$, and $\hat{b}=\sqrt{3L+2LA}b$.

This complete the prove. ■