8 Appendix A

8.1 Nomenclature

Symbol	Explanation
$\bar{L}$	The Lipschitz constant
η	Learning rate for gradient algorithm
α	Weight parameter for batch-normalization update
$\vec{x}$	The vector for one record of input data
y_j	The target output vector for the input record $\vec{x}_j$
N	Size of training set, number of all $\vec{x}$ in one epoch
D_1	Dimension of the hidden layer
$z_j^{(1)}, z_j^{(2)}$	j^{th} entry of output of the first and second hidden layer, respectively
n_1, n_2	Numbers of parameters in θ and λ, respectively
θ	Set of all trainable parameters updated by its gradient
$W^{(1)}, W^{(2)}$	$\in \theta$, weights of linear transformation between layers
$\gamma_j^{(1)}, \beta_j^{(1)}$	$\in \theta$, trainable parameters for batch-normalized output $y_j^{(1)}$
λ	Set of all batch normalization parameters determined by previous updates
μ_j	$\in \lambda$, mean of previous values of $z_j^{(1)}$
σ_j	$\in \lambda$, standard deviation of previous values of $z_j^{(1)}$
ϵ_B	The offset for batch normalization transformation

8.2 Preliminary Results

Proposition 8.1 *There exists a constant M such that, for any θ and fixed λ, we have*
$$\|\nabla \bar{f}(\theta, \lambda)\|_2^2 \leq M.$$

Proof. By Assumption 4.5, we know there exists (θ^*, λ^*) such that $\|\nabla \bar{f}(\theta^*, \lambda^*)\|_2 = 0$. Then we have

$$\begin{aligned}
&\|\nabla \bar{f}(\theta, \lambda)\|_2 \\
=&\|\nabla \bar{f}(\theta, \lambda)\|_2 - \|\nabla \bar{f}(\theta^*, \lambda^*)\|_2 \\
=&\|\nabla \bar{f}(\theta, \lambda)\|_2 - \|\nabla \bar{f}(\theta, \lambda^*)\|_2 + \|\nabla \bar{f}(\theta, \lambda^*)\|_2 - \|\nabla \bar{f}(\theta^*, \lambda^*)\|_2 \\
\leq&\|\nabla \bar{f}(\theta, \lambda) - \nabla \bar{f}(\theta, \lambda^*)\|_2 + \|\nabla \bar{f}(\theta, \lambda^*) - \nabla \bar{f}(\theta^*, \lambda^*)\|_2 \\
\leq&\sum_{i=1}^{N}\|\nabla f_i(X_i : \theta, \lambda) - \nabla f_i(X_i : \theta, \lambda^*)\|_2 + \sum_{i=1}^{N}\|\nabla f_i(X_i : \theta, \lambda^*) - \nabla f_i(X_i : \theta^*, \lambda^*)\|_2 \\
\leq&N\bar{L}(\|\lambda - \lambda^*\|_2 + \|\theta - \theta^*\|_2),
\end{aligned}$$

where the last inequality is by Assumption 4.1. We then have
$$\|\nabla \bar{f}(\theta, \lambda)\|_2^2 \leq N^2\bar{L}^2(\|\lambda - \lambda^*\|_2 + \|\theta - \theta^*\|_2)^2 \leq M,$$
because sets P and Q are compact by Assumption 4.2. □

Proposition 8.2 *We have*
$$f_i(X : \tilde{\theta}, \lambda) \leq f_i(X : \hat{\theta}, \lambda) + \nabla f_i(X : \hat{\theta}, \lambda)^T(\tilde{\theta} - \hat{\theta}) + \frac{1}{2}\bar{L}\|\tilde{\theta} - \hat{\theta}\|_2^2, \forall \tilde{\theta}, \hat{\theta}, X.$$

Proof. This is a known result of the Lipschitz-continuous condition that can be found in [5]. We have this result together with Assumption 4.1.

8.3 Proof of Theorem 4.6

Lemma 8.3 *When $\sum_{m=1}^{\infty} \alpha^{(m)} < \infty$ and $\sum_{m=1}^{\infty}\sum_{n=1}^{m} \alpha^{(m)}\eta^{(n)} < \infty$,*
$\tilde{\mu}_j^{(m)} := \dfrac{\mu_j^{(m)}}{(1-\alpha^{(1)})(1-\alpha^{(2)})...(1-\alpha^{(m)})}$ *is a Cauchy series.*

Proof. By Algorithm 1, we have

$$\mu_j^{(m)} = \alpha^{(m)} \frac{1}{N} \sum_{i=1}^{N} k^{(1)} W_{1,j,\cdot}^{(m)} X_i + (1 - \alpha^{(m)}) \mu_j^{(m-1)}. \tag{4}$$

We define $\tilde{\alpha}^{(m)} := \dfrac{\alpha^{(m)}}{(1-\alpha^{(1)})(1-\alpha^{(2)})...(1-\alpha^{(m)})}$ and $\Delta W_{1,j,\cdot}^{(m)} := W_{1,j,\cdot}^{(m)} - W_{1,j,\cdot}^{(m-1)}$. After dividing (4) by $(1-\alpha^{(1)})(1-\alpha^{(2)})...(1-\alpha^{(m)})$, we obtain

$$\tilde{\mu}_j^{(m)} = \tilde{\alpha}^{(m)} k^{(1)} \frac{1}{N} \sum_{i=1}^{N} W_{1,j,\cdot}^{(m)} X_i + \tilde{\mu}_j^{(m-1)}.$$

Then we have

$$|\tilde{\mu}_j^{(m+1)} - \tilde{\mu}_j^{(m)}| = \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} \left(|W_{1,j,\cdot}^{(m)} X_i| \right)$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} |\sum_{n=1}^{m} \Delta W_{1,j,\cdot}^{(n)} X_i| \tag{5}$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} \left| \sum_{n=1}^{m} \left(\eta^{(n)} \sum_{l=1}^{N} \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)}) \right) \cdot X_i \right|$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} \sum_{n=1}^{m} \left(\eta^{(n)} \left| \left(\sum_{l=1}^{N} \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)}) \right) \cdot X_i \right| \right)$$

$$\leq \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} \sum_{n=1}^{m} \left(\eta^{(n)} \| \sum_{l=1}^{N} \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)}) \| \cdot \|X_i\| \right) \tag{6}$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \cdot$$
$$\sum_{i=1}^{N} \sum_{n=1}^{m} \left(\eta^{(n)} \| \sum_{l=1}^{N} \left[\nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)}) - \nabla_{W_{1,j,\cdot}} \bar{f}(X_l : \theta^*, \lambda^*) \right] \| \cdot \|X_i\| \right)$$

$$\leq \tilde{\alpha}^{(m)} |k^{(1)}| \frac{1}{N} \sum_{i=1}^{N} \sum_{n=1}^{m} \eta^{(n)} (\sum_{l=1}^{N} [\| \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)}) - \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^*, \lambda^{(n)}) \|_2 +$$
$$\| \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^*, \lambda^{(n)}) - \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^*, \lambda^*) \|_2] \cdot \|X_i\|_2)$$

$$\leq \tilde{\alpha}^{(m)} |k^{(1)}| \sum_{i=1}^{N} \sum_{n=1}^{m} \eta^{(n)} \left(\bar{L} \cdot (\|W_{1,j,\cdot}^{(n)} - W_{1,j,\cdot}^*\|_2 + \|\lambda_{j,\cdot}^{(n)} - \lambda_{j,\cdot}^*\|_2) \cdot \|X_i\|_2 \right)$$

$$\leq \tilde{\alpha}^{(m)} \sum_{n=1}^{m} \left(\eta^{(n)} \right) |k^{(1)}| \sum_{i=1}^{N} \left(2\bar{L}M \|X_i\|_2 \right) \tag{7}$$

$$\leq \tilde{\alpha}^{(m)} \sum_{n=1}^{m} \eta^{(n)} \tilde{M}_{\bar{L},M}.$$

Equation (5) is due to

$$W_{1,i,j}^{(m)} = \sum_{n=1}^{m} \Delta W_{1,i,j}^{(n)}.$$

Therefore,

$$
\begin{aligned}
|\tilde{\mu}_j^{(p)} - \tilde{\mu}_j^{(q)}| &\leq \tilde{M}_{\bar{L},M} \cdot \sum_{m=p}^{q} \tilde{\alpha}^{(m)} \sum_{n=1}^{m} \eta^{(n)} \\
&= \tilde{M}_{\bar{L},M} \cdot \sum_{m=p}^{q} \tilde{\alpha}^{(m)} \sum_{n=1}^{m} \eta^{(n)} = \tilde{M}_{\bar{L},M} \cdot \sum_{m=p}^{q} \sum_{n=1}^{m} \tilde{\alpha}^{(m)} \eta^{(n)}.
\end{aligned} \tag{8}
$$

It remains to show that

$$\sum_{m=1}^{\infty} \alpha^{(m)} < \infty, \tag{9}$$

$$\sum_{m=1}^{\infty} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)} < \infty, \tag{10}$$

implies the convergence of $\{\tilde{\mu}^{(m)}\}$. By (9), we have

$$\Pi_{m=1}^{\infty}(1 - \alpha^{(m)}) > 0,$$

since

$$\ln(\Pi_{m=1}^{\infty}(1 - \alpha^{(m)})) = \sum_{m=1}^{\infty} \ln(1 - \alpha^{(m)}) > \sum_{m=1}^{\infty} -\alpha^{(m)} > -\infty.$$

It is also easy to show that there exists C and M_c such that for all $m \geq M_c$, we have

$$(1 - \alpha^{(1)})(1 - \alpha^{(2)}) \ldots (1 - \alpha^{(m)}) \geq C. \tag{11}$$

Therefore,

$$\lim_{m \to \infty} (1 - \alpha^{(1)})(1 - \alpha^{(2)}) \ldots (1 - \alpha^{(m)}) \geq C.$$

Thus the following holds:

$$\tilde{\alpha}^{(m)} \leq \frac{1}{C} \alpha^{(m)} \tag{12}$$

and

$$\sum_{m=p}^{q} \sum_{n=1}^{m} \tilde{\alpha}^{(m)} \eta^{(n)} \leq \frac{1}{C} \sum_{m=p}^{q} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)}. \tag{13}$$

From (10) and (13) it follows that the sequence $\{\tilde{\mu}_j^{(m)}\}$ is a Cauchy series. □

Lemma 8.4 *Since $\{\tilde{\mu}_j^{(m)}\}$ is a Cauchy series, $\{\mu_j^{(m)}\}$ is a Cauchy series.*

Proof. We know that

$$\mu_j^{(m)} = \tilde{\mu}_j^{(m)} (1 - \alpha^{(1)}) \ldots (1 - \alpha^{(m)}).$$

Since

$$\lim_{m \to \infty} \tilde{\mu}_j^{(m)} \to \tilde{\mu}_j$$

and

$$\lim_{m \to \infty} (1 - \alpha^{(1)}) \ldots (1 - \alpha^{(m)}) \to \tilde{C},$$

we have

$$\lim_{m \to \infty} \mu_j^{(m)} \to \tilde{\mu}_j \cdot \tilde{C}.$$

Thus $\mu_j^{(m)}$ is a Cauchy series. □

Lemma 8.5 *If $\sum_{m=1}^{\infty} \alpha^{(m)} < \infty$ and $\sum_{m=1}^{\infty} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)} < \infty$, $\{\sigma_j^{(m)}\}$ is a Cauchy series.*

Proof. We define $\sigma_j^{(m)} := \tilde{\sigma}_j^{(m)}(1-\alpha^{(1)})...(1-\alpha^{(m)})$. Then we have

$$|\tilde{\sigma}_j^{(m+1)} - \tilde{\sigma}_j^{(m)}| = \tilde{\alpha}^{(m)} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(k^{(1)}W_{1,j,\cdot}^{(m)}X_i - \mu_j^{(m)}\right)^2}$$

$$= \tilde{\alpha}^{(m)}\frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^{N}\left(k^{(1)}W_{1,j,\cdot}^{(m)}X_i - \mu_j^{(m)}\right)^2}$$

$$= \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\sqrt{\sum_{i=1}^{N}\left(W_{1,j,\cdot}^{(m)}X_i - \frac{\mu_j^{(m)}}{k^{(1)}}\right)^2}.$$

Since $\{\mu_j^{(m)}\}$ is convergent, there exists c_1, c_2 and N_1 such that for any $m > N_1$, $-\infty < c_1 < \mu_j^{(m)} < c_2 < \infty$. Therefore,

$$\begin{aligned}&|\tilde{\sigma}_j^{(m+1)} - \tilde{\sigma}_j^{(m)}| \\ \leq&\tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\max\left\{\sqrt{\sum_{i=1}^{N}\left(W_{1,j,\cdot}^{(m)}X_i - \frac{c_1}{k^{(1)}}\right)^2}, \sqrt{\sum_{i=1}^{N}\left(W_{1,j,\cdot}^{(m)}X_i - \frac{c_2}{k^{(1)}}\right)^2}\right\}.\end{aligned} \tag{14}$$

For any $\bar{C} \in \left\{\frac{c_1}{k^{(1)}}, \frac{c_2}{k^{(1)}}\right\}$, we have

$$|\tilde{\sigma}_j^{(m+1)} - \tilde{\sigma}_j^{(m)}| \leq \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(W_{1,j,\cdot}^{(m)}X_i - \bar{C}\right)^2} \tag{15}$$

$$\leq \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(|W_{1,j,\cdot}^{(m)}X_i| + |\bar{C}|\right)^2} \tag{16}$$

$$= \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(|\left(\sum_{n=1}^{m}\Delta W_{1,j,\cdot}^{(n)}\right)X_i| + |\bar{C}|\right)^2}$$

$$= \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(|\sum_{n=1}^{m}\left(\eta^{(n)}\cdot\sum_{l=1}^{N}\nabla_{W_{1,j,\cdot}}f_l(X_l:\theta^{(n)},\lambda^{(n)})\cdot X_i\right)| + |\bar{C}|\right)^2}$$

$$\leq \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(\sum_{n=1}^{m}\left(\eta^{(n)}\cdot|\sum_{l=1}^{N}\nabla_{W_{1,j,\cdot}}f_l(X_l:\theta^{(n)},\lambda^{(n)})\cdot X_i|\right) + |\bar{C}|\right)^2} \tag{17}$$

$$\leq \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(\sum_{n=1}^{m}\eta^{(n)}\|\sum_{l=1}^{N}\nabla_{W_{1,j,\cdot}}f_l(X_l:\theta^{(n)},\lambda^{(n)})\|\cdot\|X_i\| + |\bar{C}|\right)^2} \tag{18}$$

$$\leq \tilde{\alpha}^{(m)}\frac{|k^{(1)}|}{\sqrt{N}}\cdot\sqrt{\sum_{i=1}^{N}\left(\sum_{n=1}^{m}\eta^{(n)}\left(2N\bar{L}M\|X_i\|_2\right) + |\bar{C}|\right)^2} \tag{19}$$

$$\leq \tilde{\alpha}^{(m)} \frac{|k^{(1)}|}{\sqrt{N}} \cdot \sqrt{N \cdot \left(\tilde{M}_{\bar{L},M} \sum_{n=1}^{m} \eta^{(n)} + |\bar{C}| \right)^2} \tag{20}$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \cdot \sqrt{\left(\tilde{M}_{\bar{L},M} \sum_{n=1}^{m} \eta^{(n)} + |\bar{C}| \right)^2}$$

$$= \tilde{\alpha}^{(m)} |k^{(1)}| \cdot \left(\tilde{M}_{\bar{L},M} \sum_{n=1}^{m} \eta^{(n)} + |\bar{C}| \right). \tag{21}$$

Inequality (15) is by plugging $\bar{C} \in \left\{ \frac{c_1}{k^{(1)}}, \frac{c_2}{k^{(1)}} \right\}$ into (14). Inequality (16) is by the following fact:

$$\sqrt{\sum_{i=1}^{n} (a_i - c)^2} \leq \max \left\{ \sqrt{\sum_{i=1}^{n} (|a_i| - c)^2}, \sqrt{\sum_{i=1}^{n} (|a_i| + c)^2} \right\} = \sqrt{\sum_{i=1}^{n} (|a_i| + |c|)^2}, \tag{22}$$

where b and a_i for every i are arbitrary real scalars. Besides, (22) is due to

$$-2a_i c \leq \max\{-2|a_i|c, 2|a_i|c\}.$$

Inequalities (17), (18) and (19) follow from the square function being increasing for nonnegative numbers. Besides these facts, (19) is also by the same techniques we used in (6)-(7) where we bound the derivatives with the Lipschitz continuity in the following inequality:

$$\|\sum_{l=1}^{N} \nabla_{W_{1,j,\cdot}} f_l(X_l : \theta^{(n)}, \lambda^{(n)})\| \leq 2N\bar{L}M.$$

Inequality (20) is by collecting the bounded terms into a single bound $\tilde{M}_{\bar{L},M}$. Therefore,

$$|\tilde{\sigma}_j^{(q)} - \tilde{\sigma}_j^{(p)}| \leq \sum_{m=p}^{q-1} |\tilde{\sigma}_j^{(m+1)} - \tilde{\sigma}_j^{(m)}| \leq \sum_{m=p}^{q-1} \tilde{\alpha}^{(m)} |k^{(1)}| \cdot \left(\tilde{M}_{\bar{L},M} \sum_{n=1}^{m} \eta^{(n)} + |\bar{C}| \right). \tag{23}$$

Using the similar methods in deriving (9) and (10), it can be seen that a set of sufficient conditions ensuring the convergence for $\{\tilde{\sigma}_j^{(m)}\}$ is:

$$\sum_{m=1}^{\infty} \alpha^{(m)} < \infty,$$
$$\sum_{m=1}^{\infty} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)} < \infty.$$

Therefore, the convergence conditions for $\{\sigma_j^{(m)}\}$ are the same as for $\{\mu_j^{(m)}\}$. □

It is clear that these lemmas establish the proof of Theorem 4.6.

8.4 Consequences of Theorem 4.6

Proposition 8.6 *Under the assumptions of Theorem 4.6, we have*

$$|\lambda^{(m)} - \bar{\lambda}|_\infty \leq a_m,$$

where

$$a_m = M_1 \sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \eta^{(j)} + M_2 \sum_{i=m}^{\infty} \alpha^{(i)} \tag{24}$$

and M_1 and M_2 are constants.

Proof. For the upper bound of $\sigma_j^{(m)}$, by (21), we have

$$|\tilde{\sigma}_j^{(q)} - \tilde{\sigma}_j^{(p)}| \leq \sum_{m=p}^{q-1} \tilde{\alpha}^{(m)} |k^{(1)}| \left(\tilde{M}_{\bar{L},M} \sum_{n=1}^{m} \eta^{(n)} + |\bar{C}| \right).$$

We define $\tilde{\bar{\sigma}}_j := \dfrac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(u)})...}$. Therefore,

$$\begin{aligned} |\tilde{\bar{\sigma}}_j - \tilde{\sigma}_j^{(m)}| &\leq \sum_{i=m}^{\infty} \tilde{\alpha}^{(i)} |k^{(1)}| \left(\tilde{M}_{\bar{L},M} \sum_{j=1}^{i} \eta^{(j)} + |\bar{C}| \right) \\ &\leq \frac{|k^{(1)}|}{C} \sum_{i=m}^{\infty} \alpha^{(i)} \left(\tilde{M}_{\bar{L},M} \sum_{j=1}^{i} \eta^{(j)} + |\bar{C}| \right). \end{aligned} \tag{25}$$

The first inequality comes by substituting p by m and by taking $\lim$ as $q \to \infty$ in (23). The second inequality comes from (11). We then obtain,

$$\begin{aligned} &\left|\sigma_j^{(m)} - \bar{\sigma}_j\right| \\ =&(1-\alpha^{(1)})...(1-\alpha^{(m)}) \left| \frac{\sigma_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} \right| \\ \leq&(1-\alpha^{(1)})...(1-\alpha^{(m)}) [\left| \frac{\sigma_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(u)})...} \right| + \\ &\qquad \left| \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(u)})...} \right|] \\ \leq& \left| \tilde{\sigma}_j^{(m)} - \tilde{\sigma}_j^{(\infty)} \right| + \left| \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\sigma}_j^{(\infty)} \right| \\ =& \left| \tilde{\sigma}_j^{(m)} - \tilde{\sigma}_j^{(\infty)} \right| + \left| \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\sigma}_j}{(1-\alpha^{(1)})...(1-\alpha^{(u)})...} \right| \\ =& \left| \tilde{\sigma}_j^{(m)} - \tilde{\sigma}_j^{(\infty)} \right| + \bar{\sigma}_j \left| \frac{(1-\alpha^{(m+1)})...(1-\alpha^{(u)})... - 1}{(1-\alpha^{(1)})...(1-\alpha^{(u)})...} \right| \\ \leq& \left| \tilde{\sigma}_j^{(m)} - \tilde{\sigma}_j^{(\infty)} \right| + \frac{\bar{\sigma}_j}{C} |1 - (1-\alpha^{(m+1)})...(1-\alpha^{(u)})...| \\ \leq& \left| \tilde{\sigma}_j^{(m)} - \tilde{\sigma}_j^{(\infty)} \right| + \frac{\bar{\sigma}_j}{C} \sum_{n=m+1}^{\infty} \alpha^{(n)}. \end{aligned} \tag{26}$$

The second inequality is by $(1-\alpha^{(1)})...(1-\alpha^{(m)}) < 1$, the third inequality is by (11) and the last inequality can be easily seen by induction. By (26), we obtain

$$|\bar{\sigma}_j - \sigma_j^{(m)}| = \lim_{M \to \infty} |\sigma_j^{(M)} - \sigma_j^{(m)}| \leq |\tilde{\sigma}_j - \tilde{\sigma}_j^{(m)}| + \frac{\bar{\sigma}_j}{C} \sum_{n=m+1}^{\infty} \alpha^{(n)}. \tag{27}$$

Therefore, we have

$$
\begin{aligned}
&|\bar{\sigma}_j - \sigma_j^{(m)}| \\
\leq &|\bar{\sigma}_j - \tilde{\sigma}_j^{(m)}| + \frac{\bar{\sigma}_j}{C} \sum_{n=m+1}^{\infty} \alpha^{(n)} \\
\leq &\sum_{i=m}^{\infty} \tilde{\alpha}^{(i)} |k^{(1)}| \cdot \left(\tilde{M}_{\bar{L},M} \sum_{j=1}^{i} \eta^{(j)} + |\bar{C}| \right) + \frac{\bar{\sigma}_j}{C} \sum_{i=m+1}^{\infty} \alpha^{(i)} \\
\leq &\sum_{i=m}^{\infty} \frac{1}{C} \alpha^{(i)} |k^{(1)}| \cdot \left(\tilde{M}_{\bar{L},M} \sum_{j=1}^{i} \eta^{(j)} + |\bar{C}| \right) + \frac{\bar{\sigma}_j}{C} \sum_{i=m+1}^{\infty} \alpha^{(i)} \\
\leq &\sum_{i=m}^{\infty} \frac{1}{C} \alpha^{(i)} |k^{(1)}| \cdot \left(\tilde{M}_{\bar{L},M} \sum_{j=1}^{i} \eta^{(j)} + |\bar{C}| \right) + \frac{\bar{\sigma}_j}{C} \sum_{i=m}^{\infty} \alpha^{(i)} \\
= &\frac{\tilde{M}_{\bar{L},M} |k^{(1)}|}{C} \sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \eta^{(j)} + \left(\frac{\bar{\sigma}_j}{C} + \frac{|k^{(1)}||\bar{C}|}{C} \right) \sum_{i=m}^{\infty} \alpha^{(i)}.
\end{aligned}
\tag{28}
$$

The first inequality is by (27), the second inequality is by (23), the third inequality is by (12) and the fourth inequality is by adding the nonnegative term $\frac{\bar{\sigma}_j}{C}\alpha^{(m)}$ to the right-hand side.

For the upper bound of $\mu_j^{(m)}$, we have

$$
\begin{aligned}
&\left| \mu_j^{(m)} - \bar{\mu}_j \right| \\
= &(1-\alpha^{(1)})...(1-\alpha^{(m)}) \left| \frac{\mu_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} \right| \\
\leq &(1-\alpha^{(1)})...(1-\alpha^{(m)}) [\left| \frac{\mu_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(\infty)})} \right| + \\
&\left| \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(\infty)})} \right|] \\
\leq &\left| \tilde{\mu}^{(m)} - \tilde{\mu}^{(\infty)} \right| + \left| \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\mu}^{(\infty)} \right|.
\end{aligned}
\tag{29}
$$

Let us define $A_m := \left| \tilde{\mu}^{(m)} - \tilde{\mu}^{(\infty)} \right|$ and $B_m := \left| \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\mu}^{(\infty)} \right|$. Recall from Theorem 4.6 that $\{\mu_j^{(m)}\}$ is a Cauchy series, by (8),

$$
|\tilde{\mu}_j^{(p)} - \tilde{\mu}_j^{(q)}| \leq \bar{M}_{\bar{L},M} \cdot \sum_{m=p}^{q} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)}.
$$

Therefore, the first term in (29) is bounded by

$$
|\tilde{\mu}_j^{(m)} - \tilde{\mu}_j^{\infty}| \leq \tilde{M}_{\bar{L},M} \cdot \sum_{i=m}^{\infty} \sum_{n=1}^{i} \alpha^{(i)} \eta^{(n)} < \infty. \tag{30}
$$

For the second term in (29), recall that $C := (1-\alpha^{(1)})...(1-\alpha^{(u)})...$. Then we have

$$
\begin{aligned}
&C \cdot \left| \frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\mu}^{(\infty)} \right| \\
= &\bar{\mu}_j |1 - (1-\alpha^{(m+1)})...(1-\alpha^{(u)})...| \\
\leq &\bar{\mu}_j \sum_{i=m+1}^{\infty} \alpha^{(i)},
\end{aligned}
$$

where the last inequality can be easily seen by induction. Therefore, the second term in (29) is bounded by

$$\left|\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\mu}^{(\infty)}\right| \leq \frac{\bar{\mu}_j}{C}\sum_{i=m+1}^{\infty}\alpha^{(i)}. \tag{31}$$

From these we obtain

$$\begin{aligned}
&\left|\mu_j^{(m)} - \bar{\mu}_j\ \right| \\
\leq &\left|\tilde{\mu}^{(m)} - \tilde{\mu}^{(\infty)}\right| + \left|\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})} - \tilde{\mu}^{(\infty)}\right| \\
\leq &\tilde{M}_{\bar{L},M}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} + \frac{\bar{\mu}_j}{C}\sum_{i=m+1}^{\infty}\alpha^{(i)}.
\end{aligned} \tag{32}$$

The first inequality is by (29) and the second inequality is by (30) and (31). Combining (28) and (32), we have that

$$|\lambda^{(m)} - \bar{\lambda}|_\infty = \max(|\mu^{(m)} - \bar{\mu}|_\infty, |\sigma^{(m)} - \bar{\sigma}|_\infty) \leq M_1\sum_{i=m}^{\infty}\sum_{j=1}^{i}\alpha^{(i)}\eta^{(j)} + M_2\sum_{i=m}^{\infty}\alpha^{(i)},$$

where M_1 and M_2 are constants defined as

$$M_1 = \max(\frac{\tilde{M}_{\bar{L},M}|k^{(1)}|}{C}, \bar{M}_{\bar{L},M})$$

and

$$M_2 = \max(\frac{\bar{\sigma}_j + |k^{(1)}||\bar{C}|}{C}, \frac{\bar{\mu}_j}{C}).\square$$

Proposition 8.7 *Under the assumptions of Theorem 4.6,*

$$-\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})^T \cdot \nabla\bar{f}(\theta^{(m)}, \lambda^{(m)}) \leq -\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|^2 + \bar{L}M\sqrt{n_2}a_m,$$

where a_m is defined in Proposition 8.6.

Proof. For simplicity of the proof, let us define

$$x^{(m)} := \nabla\bar{f}(\theta^{(m)}, \bar{\lambda}), \quad y^{(m)} := \nabla\bar{f}(\theta^{(m)}, \lambda^{(m)}).$$

We have

$$|x^{(m)} - y^{(m)}|_\infty \leq \|x^{(m)} - y^{(m)}\|_2 \leq \bar{L}\|\lambda^{(m)} - \bar{\lambda}\|_2 \leq \bar{L}\sqrt{n_2}\|\lambda^{(m)} - \bar{\lambda}\|_\infty \leq \bar{L}\sqrt{n_2}a_m, \tag{33}$$

where $\sqrt{n_2}$ is the dimension of λ. The second inequality is by Assumption 4.1 and the fourth inequality is by Proposition 8.6. Inequality (33) implies that for all m and i, we have

$$|x_i^{(m)} - y_i^{(m)}| \leq \bar{L}\sqrt{n_2}a_m.$$

It remains to show

$$-\sum_i y_i^{(m)}x_i^{(m)} \leq -\sum_i {x_i^{(m)}}^2 + \bar{L}M\sqrt{n_2}a_m, \forall i, m. \tag{34}$$

This is established by the following four cases.

1) If $x_i^{(m)} \geq 0, x_i^{(m)} - y_i^{(m)} \geq 0$, then $x_i^{(m)} \leq \bar{L}\sqrt{n_2}a_m + y_i^{(m)}$. Thus $-x_i^{(m)}y_i^{(m)} \leq -{x_i^{(m)}}^2 + \bar{L}M\sqrt{n_2}a_m$ by Proposition 8.1.

2) If $x_i^{(m)} \geq 0, x_i^{(m)} - y_i^{(m)} \leq 0$, then $x_i^{(m)} \leq y_i^{(m)}$, ${x_i^{(m)}}^2 \leq x_i^{(m)} \cdot y_i^{(m)}$ and $-x_i^{(m)}y_i^{(m)} \leq -{x_i^{(m)}}^2$.

3) If $x_i^{(m)} < 0, x_i^{(m)} - y_i^{(m)} \geq 0$, then $x_i^{(m)} \geq y_i^{(m)}$, ${x_i^{(m)}}^2 \leq x_i^{(m)} \cdot y_i^{(m)}$ and $-x_i^{(m)}y_i^{(m)} \leq -{x_i^{(m)}}^2$.

4) If $x_i^{(m)} < 0, x_i^{(m)} - y_i^{(m)} \leq 0$, then $y_i^{(m)} - x_i^{(m)} \leq \bar{L}\sqrt{n_2}a_m, y_i^{(m)}x_i^{(m)} - {x_i^{(m)}}^2 \geq \bar{L}\sqrt{n_2}a_m x_i^{(m)}$ and $-y_i^{(m)}x_i^{(m)} \leq -{x_i^{(m)}}^2 - \bar{L}\sqrt{n_2}a_m x_i^{(m)} \leq -{x_i^{(m)}}^2 + \bar{L}M\sqrt{n_2}a_m$. The last inequality is by Proposition 8.1.

All these four cases yield (34). $\square$

Proposition 8.8 *Under the assumptions of Theorem 4.6, we have*

$$\bar{f}(\theta^{(m+1)}, \bar{\lambda}) \leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - \eta^{(m)}\|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 + \eta^{(m)}\bar{L}M\sqrt{n_2}a_m + \frac{1}{2}(\eta^{(m)})^2 \cdot N\bar{L}M,$$

where M is a constant and a_m is defined in Proposition 8.6.

Proof. By Proposition 8.2,

$$f_i(X_i : \tilde{\theta}, \lambda) \leq f_i(X_i : \hat{\theta}, \lambda) + \nabla f_i(X_i : \hat{\theta}, \lambda)^T(\tilde{\theta} - \hat{\theta}) + \frac{1}{2}\bar{L}\|\tilde{\theta} - \hat{\theta}\|_2^2.$$

Therefore, we can sum it over the entire training set from $i = 1$ to N to obtain

$$\bar{f}(\tilde{\theta}, \lambda) \leq \bar{f}(\hat{\theta}, \lambda) + \nabla \bar{f}(\hat{\theta}, \lambda)^T(\tilde{\theta} - \hat{\theta}) + \frac{N}{2}\bar{L}\|\tilde{\theta} - \hat{\theta}\|_2^2. \tag{35}$$

In Algorithm 1, we define the update of θ in the following full gradient way:

$$\theta^{(m+1)} := \theta^{(m)} - \eta^{(m)} \cdot \sum_{i=1}^{N} \cdot \nabla f_i(X_i : \theta^{(m)}, \lambda^{(m)}),$$

which implies

$$\theta^{(m+1)} - \theta^{(m)} = -\eta^{(m)} \cdot \nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}). \tag{36}$$

By (36) we have $\tilde{\theta} - \hat{\theta} = \theta^{(m+1)} - \theta^{(m)} = -\eta^{(m)}\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})$. We now substitute $\tilde{\theta} := \theta^{(m+1)}$, $\hat{\theta} := \theta^{(m)}$ and $\lambda := \bar{\lambda}$ into (35) to obtain

$$\begin{aligned}
&\bar{f}(\theta^{(m+1)}, \bar{\lambda}) \\
&\leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - \eta^{(m)}\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})^T \nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}) + (\eta^{(m)})^2 \cdot \frac{N\bar{L}}{2}\|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})\|_2^2 \\
&\leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - \eta^{(m)}\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})^T \nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}) + (\eta^{(m)})^2 \cdot \frac{N\bar{L}M}{2} \\
&\leq \bar{f}(\theta^{(m)}, \bar{\lambda}) + \eta^{(m)}\left(-\|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 + \bar{L}M\sqrt{n_2}a_m\right) + \frac{1}{2}(\eta^{(m)})^2 \cdot N\bar{L}M \\
&= \bar{f}(\theta^{(m)}, \bar{\lambda}) - \eta^{(m)}\|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 + \eta^{(m)}\bar{L}M\sqrt{n_2}a_m + \frac{1}{2}(\eta^{(m)})^2 \cdot N\bar{L}M.
\end{aligned} \tag{37}$$

The first inequality is by plugging (36) into (35), the second inequality comes from Proposition 8.1 and the third inequality comes from Proposition 8.7. $\square$

8.5 Proof of Theorem 4.10

Here we show Theorem 4.10 as the consequence of Theorem 4.6 and Lemmas 4.7, 4.8 and 4.9.

8.5.1 Proof of Lemma 4.7

Here we show Lemma 4.7 as the consequence of Lemmas 8.9, 8.10 and 8.11.

Lemma 8.9

$$\sum_{m=1}^{\infty} \sum_{i=m}^{\infty} \sum_{n=1}^{i} \alpha^{(i)}\eta^{(n)} < \infty$$

and

$$\sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} < \infty$$

is a set of sufficient condition to ensure

$$\sum_{m=1}^{\infty}|\bar{\sigma}_j - \sigma_j^{(m)}| < \infty, \forall j. \tag{38}$$

Proof. By plugging (27) and (25) into (38), we have the following for all j:

$$\begin{aligned}
&\sum_{m=1}^{\infty}\left|\bar{\sigma}_j - \sigma_j^{(m)}\right| \\
\leq &\sum_{m=1}^{\infty}\left(|\bar{\sigma}_j - \tilde{\sigma}_j^{(m)}| + \frac{\bar{\sigma}_j}{C}\sum_{n=m+1}^{\infty}\alpha^{(n)}\right) \\
\leq &\sum_{m=1}^{\infty}\left[\frac{|k^{(1)}|}{C}\sum_{i=m}^{\infty}\left[\alpha^{(i)}\left(\tilde{M}_{\bar{L},M}\sum_{j=1}^{i}\eta^{(j)} + |\bar{C}|\right)\right] + \frac{\bar{\sigma}_j}{C}\sum_{n=m+1}^{\infty}\alpha^{(n)}\right] \\
\leq &\frac{|k^{(1)}|\cdot\tilde{M}_{\bar{L},M}}{C}\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\alpha^{(i)}\sum_{j=1}^{i}\eta^{(j)} + \frac{\bar{\sigma}_j + |k^{(1)}||\bar{C}|}{C}\sum_{m=1}^{\infty}\sum_{n=m+1}^{\infty}\alpha^{(n)}.
\end{aligned} \tag{39}$$

It is easy to see that the the following conditions are sufficient for right-hand side of (39) to be finite:

$$\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} < \infty$$

and

$$\sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} < \infty.$$

Therefore, we obtain

$$\sum_{m=1}^{\infty}|\bar{\sigma}_j - \sigma_j^{(m)}| < \infty, \forall j. \qquad \square$$

Lemma 8.10 *Under Assumption 4.4,*

$$\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} < \infty$$

is a set of sufficient conditions to ensure

$$\limsup_{M\to\infty}\sum_{m=1}^{M}\left|\bar{f}(\theta^{(m)},\lambda^{(m)}) - \bar{f}(\theta^{(m)},\bar{\lambda})\right| < \infty.$$

Proof. By Assumption 4.4, we have

$$\|l_i(x) - l_i(y)\| \leq \hat{M}\|x-y\| \leq \hat{M}\sum_{i=1}^{D}|x_i - y_i|. \tag{40}$$

By the definition of $f_i(\cdot)$, we then have

$$\begin{aligned}
&\sum_{m=1}^{\infty}\left|\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{(m)},\bar{\lambda})\right| \\
=&\sum_{m=1}^{\infty}\left|\sum_{i=1}^{N}\left(l_i(X_i:\theta^{(m)},\lambda^{(m)})+c_2\|\theta^{(m)}\|_2^2\right)-\sum_{i=1}^{N}\left(l_i(X_i:\theta^{(m)},\bar{\lambda})+c_2\|\theta^{(m)}\|_2^2\right)\right| \\
=&\sum_{m=1}^{\infty}\left|\sum_{i=1}^{N}\left(l_i(X_i:\theta^{(m)},\lambda^{(m)})-l_i(X_i:\theta^{(m)},\bar{\lambda})\right)\right| \\
\leq&\sum_{m=1}^{\infty}\sum_{i=1}^{N}\left|\left(l_i(X_i:\theta^{(m)},\lambda^{(m)})-l_i(X_i:\theta^{(m)},\bar{\lambda})\right)\right| \\
\leq& M_2\sum_{m=1}^{\infty}\sum_{j=1}^{D}\sum_{i=1}^{N}\left|\frac{k^{(1)}W_{1,j,\cdot}^{(m)}X_i-\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}-\frac{k^{(1)}W_{1,j,\cdot}^{(m)}X_i-\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}\right| \\
=& M_2\sum_{m=1}^{\infty}\sum_{j=1}^{D}\sum_{i=1}^{N}\left(\left|(k^{(1)}W_{1,j,\cdot}^{(m)}X_i)\left(\frac{1}{\sigma_j^{(m)}+\epsilon_B}-\frac{1}{\bar{\sigma}_j+\epsilon_B}\right)+\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right) \\
\leq& M_2\sum_{m=1}^{\infty}\sum_{j=1}^{D}\sum_{i=1}^{N}\left(\left|(k^{(1)}W_{1,j,\cdot}^{(m)}X_i)\left(\frac{1}{\sigma_j^{(m)}+\epsilon_B}-\frac{1}{\bar{\sigma}_j+\epsilon_B}\right)\right|+\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right) \\
\leq& M_3\sum_{m=1}^{\infty}\sum_{j=1}^{D}\left(\sum_{i=1}^{N}|(k^{(1)}W_{1,j,\cdot}^{(m)}X_i)|\left|\frac{1}{\sigma_j^{(m)}+\epsilon_B}-\frac{1}{\bar{\sigma}_j+\epsilon_B}\right|+\sum_{i=1}^{N}\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right) \\
=& M_3\sum_{m=1}^{\infty}\sum_{j=1}^{D}\left(\sum_{i=1}^{N}|(k^{(1)}W_{1,j,\cdot}^{(m)}X_i)|\left|\frac{\bar{\sigma}_j-\sigma_j^{(m)}}{(\sigma_j^{(m)}+\epsilon_B)(\bar{\sigma}_j+\epsilon_B)}\right|+N\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right) \\
\leq& M_3\sum_{m=1}^{\infty}\sum_{j=1}^{D}\left(\sum_{i=1}^{N}|k^{(1)}||W_{1,j,\cdot}^{(m)}X_i|\left|\frac{\bar{\sigma}_j-\sigma_j^{(m)}}{\epsilon_B^2}\right|+N\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right). \qquad (41)
\end{aligned}$$

The first inequality is by the Cauchy-Schwarz inequality, and the second one is by (40). To show the finiteness of (41), we only need to show the following two statements:

$$\sum_{m=1}^{\infty}\sum_{i=1}^{N}|k^{(1)}||W_{1,j,\cdot}^{(m)}X_i|\left|\frac{\bar{\sigma}_j-\sigma_j^{(m)}}{\epsilon_B^2}\right|<\infty,\forall j \qquad (42)$$

and

$$\sum_{m=1}^{\infty}\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|<\infty,\forall j. \qquad (43)$$

Proof of (42)*:* For all j we have

$$\begin{aligned}
&\sum_{m=1}^{\infty}\sum_{i=1}^{N}|k^{(1)}||W_{1,j,\cdot}^{(m)}X_i|\left|\frac{\bar{\sigma}_j-\sigma_j^{(m)}}{\epsilon_B^2}\right| \\
\leq&\sum_{m=1}^{\infty}|k^{(1)}|NDM\max_i\|X_i\|\frac{1}{\epsilon_B^2}\left|\bar{\sigma}_j-\sigma_j^{(m)}\right| \qquad (44) \\
=&|k^{(1)}|NDM\max_i\|X_i\|\frac{1}{\epsilon_B^2}\sum_{m=1}^{\infty}\left|\bar{\sigma}_j-\sigma_j^{(m)}\right|.
\end{aligned}$$

The inequality comes from $|W_{1,j,\cdot}^{(m)}X_i|\leq DM\|X_i\|_2$, where D is the dimension of X_i and M is the element-wise upper bound for $W_{1,j,\cdot}^{(m)}$ in Assumption 4.2.

Finally, we invoke Lemma 8.3 to assert that $\sum_{m=1}^{\infty} \left|\bar{\sigma}_j - \sigma_j^{(m)}\right|$ is finite.

Proof of (43)*:* For all j we have

$$\begin{aligned}
&\sum_{m=1}^{\infty} \left| \frac{\bar{\mu}_j}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| \\
\leq &\sum_{m=1}^{\infty} \left| \frac{\bar{\mu}_j}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} \right| + \sum_{m=1}^{\infty} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right|.
\end{aligned} \tag{45}$$

The first term in (45) is finite since $\{\mu_j^{(m)}\}$ is a Cauchy series. For the second term, we know that there exists a constant M such that for all $m \geq M$, $\mu_j^{(m)} \leq \bar{\mu} + 1$. This is also by the fact that $\{\mu_j^{(m)}\}$ is a Cauchy series and it converges to $\bar{\mu}$. Therefore, the second term in (45) becomes

$$\begin{aligned}
&\sum_{m=1}^{M-1} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| + \sum_{m=M}^{\infty} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| \\
\leq &\sum_{m=1}^{M-1} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| + \sum_{m=M}^{\infty} (\bar{\mu} + 1) \left| \frac{1}{\bar{\sigma}_j + \epsilon_B} - \frac{1}{\sigma_j^{(m)} + \epsilon_B} \right|.
\end{aligned} \tag{46}$$

Noted that function $f(\sigma) = \dfrac{1}{\sigma + \epsilon_B}$ is Lipschitz continuous since its gradient is bounded by $\dfrac{1}{\epsilon_B^2}$. Therefore we can choose $\dfrac{1}{\epsilon_B^2}$ as the Lipschitz constant for $f(\sigma)$. We then have the following inequality:

$$\left| \frac{1}{\bar{\sigma}_j + \epsilon_B} - \frac{1}{\sigma_j^{(m)} + \epsilon_B} \right| \leq \frac{1}{\epsilon_B^2} |\bar{\sigma}_j - \sigma_j^{(m)}|. \tag{47}$$

Plugging (47) into (46), we obtain

$$\begin{aligned}
&\sum_{m=1}^{M-1} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| + \sum_{m=M}^{\infty} (\bar{\mu} + 1) \left| \frac{1}{\bar{\sigma}_j + \epsilon_B} - \frac{1}{\sigma_j^{(m)} + \epsilon_B} \right| \\
\leq &\sum_{m=1}^{M-1} \left| \frac{\mu_j^{(m)}}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| + \sum_{m=M}^{\infty} \frac{(\bar{\mu} + 1)}{\epsilon_B^2} |\bar{\sigma}_j - \sigma_j^{(m)}|,
\end{aligned}$$

where the first term is finite by the fact that M is a finite constant. We have shown the condition for the second term to be finite in Lemma 8.9. Therefore,

$$\sum_{m=1}^{\infty} \left| \frac{\bar{\mu}_j}{\bar{\sigma}_j + \epsilon_B} - \frac{\mu_j^{(m)}}{\sigma_j^{(m)} + \epsilon_B} \right| < \infty, \forall j.$$

By (42) and (43), we have that the right-hand side of (41) is finite. It means that the left-hand side of (41) is finite. Thus,

$$\sum_{m=1}^{\infty} \left| \bar{f}(\theta^{(m)}, \lambda^{(m)}) - \bar{f}(\theta^{(m)}, \bar{\lambda}) \right| < \infty. \qquad \square$$

Lemma 8.11 *If*

$$\sum_{m=1}^{\infty} \sum_{i=m}^{\infty} \sum_{n=1}^{i} \alpha^{(i)} \eta^{(n)} < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} \alpha^{(n)} < \infty,$$

then

$$\limsup_{M \to \infty} \sum_{m=1}^{M} \eta^{(m)} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 < \infty.$$

Proof. For simplicity of the proof, we define

$$T^{(M)} := \sum_{m=1}^{M} \eta^{(m)} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2,$$

$$O^{(m)} := \bar{f}(\theta^{(m+1)}, \lambda^{(m+1)}) - \bar{f}(\theta^{(m)}, \lambda^{(m)}),$$

$$\Delta_1^{(m+1)} := \bar{f}(\theta^{(m+1)}, \lambda^{(m+1)}) - \bar{f}(\theta^{(m+1)}, \bar{\lambda}),$$

$$\Delta_2^{(m)} := \bar{f}(\theta^{(m+1)}, \bar{\lambda}) - \bar{f}(\theta^{(m)}, \bar{\lambda}),$$

where $\bar{\lambda}$ is the converged value of λ in Theorem 4.6. Therefore,

$$O^{(m)} = \Delta_1^{(m+1)} + \Delta_1^{(m)} + \Delta_2^{(m)} \leq |\Delta_1^{(m+1)}| + |\Delta_1^{(m)}| + \Delta_2^{(m)}. \tag{48}$$

By Proposition 8.8,

$$\Delta_2^{(m)} \leq \quad -\eta^{(m)} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 + \eta^{(m)} \bar{L} M \sqrt{n_2} a_m + \frac{1}{2} (\eta^{(m)})^2 \cdot N \bar{L} M. \tag{49}$$

We sum the inequality (48) from 1 to K with respect to m and plug (49) into it to obtain

$$\begin{aligned}
\sum_{m=1}^{K} O^{(m)} \leq & \sum_{m=1}^{K} |\Delta_1^{(m+1)}| + \sum_{m=1}^{K} |\Delta_1^{(m)}| - \sum_{m=1}^{K} \{\eta^{(m)} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2\} \\
& + \sum_{m=1}^{K} \eta^{(m)} \bar{L} M \sqrt{n_2} a_m + \sum_{m=1}^{K} \{\frac{1}{2} (\eta^{(m)})^2 N \bar{L} M\} \\
= & \sum_{m=1}^{K} |\Delta_1^{(m+1)}| + \sum_{m=1}^{K} |\Delta_1^{(m)}| - T^{(K)} \\
& + \bar{L}^2 \sqrt{n_2} \cdot \sum_{m=1}^{K} \eta^{(m)} a_m + \sum_{m=1}^{K} \{\frac{1}{2} (\eta^{(m)})^2 N \bar{L} M\}.
\end{aligned}$$

From this, we have:

$$\begin{aligned}
\limsup_{K \to \infty} T^{(K)} \leq & \limsup_{K \to \infty} \frac{-1}{c_1} (\bar{f}(\theta^{(K)}, \lambda^{(K)}) - \bar{f}(\theta^{(1)}, \lambda^{(1)})) \\
& + \limsup_{K \to \infty} \frac{1}{c_1} \sum_{m=1}^{K} (|\Delta_1^{(m+1)}| + |\Delta_1^{(m)}|) \\
& + \limsup_{K \to \infty} \bar{L}^2 \sqrt{n_2} \sum_{m=1}^{K} \eta^{(m)} a_m \\
& + \limsup_{K \to \infty} \frac{N \bar{L} K}{2 c_1} \sum_{m=1}^{K} {\eta^{(m)}}^2.
\end{aligned} \tag{50}$$

Next we show that each of the four terms in the right-hand side of (50) is finite, respectively. For the first term,

$$\limsup_{K \to \infty} \frac{-1}{c_1} (\bar{f}(\theta^{(K)}, \lambda^{(K)}) - \bar{f}(\theta^{(1)}, \lambda^{(1)})) < \infty$$

is by the fact that the parameters $\{\theta, \lambda\}$ are in compact sets, which implies that the image of $f_i(\cdot)$ is in a bounded set.

For the second term, we showed its finiteness in Lemma 8.10.

For the third term, by (24), we have

$$\begin{aligned}
&\limsup_{K\to\infty}\sum_{m=1}^{K}\eta^{(m)}a_m\\
=&\limsup_{K\to\infty}\sum_{m=1}^{K}\eta^{(m)}\left(K_1\sum_{i=m}^{\infty}\sum_{j=1}^{i}\alpha^{(i)}\eta^{(j)}+K_2\sum_{i=m}^{\infty}\alpha^{(i)}\right)\\
=&K_1\limsup_{K\to\infty}\sum_{m=1}^{K}\eta^{(m)}\left(\sum_{i=m}^{\infty}\sum_{j=1}^{i}\alpha^{(i)}\eta^{(j)}\right)+K_2\limsup_{K\to\infty}\sum_{m=1}^{K}\eta^{(m)}\sum_{i=m}^{\infty}\alpha^{(i)}.
\end{aligned} \tag{51}$$

The right-hand side of (51) is finite because

$$\sum_{m=1}^{\infty}\eta^{(m)}\left(\sum_{i=m}^{\infty}\sum_{j=1}^{i}\alpha^{(i)}\eta^{(j)}\right)<\sum_{m=1}^{\infty}\left(\sum_{i=m}^{\infty}\sum_{j=1}^{i}\alpha^{(i)}\eta^{(j)}\right)<\infty \tag{52}$$

and

$$\sum_{m=1}^{\infty}\eta^{(m)}\sum_{i=m}^{\infty}\alpha^{(i)}<\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\alpha^{(i)}<\infty. \tag{53}$$

The second inequalities in (52) and (53) come from the stated assumptions of this lemma.

For the fourth term,

$$\limsup_{K\to\infty}\frac{N\bar{L}M}{2c}\sum_{m=1}^{K}{\eta^{(m)}}^2<\infty$$

holds, because we have $\sum_{m=1}^{\infty}(\eta^{(m)})^2<\infty$ in Assumption 4.3. Therefore, $T^{(\infty)}=\sum_{m=1}^{\infty}\eta^{(m)}\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2^2<\infty$ holds. □

In Lemmas 8.9, 8.10 and 8.11, we show that $\{\sigma^{(m)}\}$ and $\{\mu^{(m)}\}$ are Cauchy series, hence Lemma 4.7 holds.

8.5.2 Proof of Lemma 4.8

This proof is similar to the the proof by Bertsekas and Tsitsiklis [4].

Proof. By Theorem 4.7, we have

$$\limsup_{M\to\infty}\sum_{m=1}^{M}\eta^{(m)}\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2^2<\infty. \tag{54}$$

If there exists a $\epsilon>0$ and an integer $\bar{m}$ such that

$$\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2\geq\epsilon$$

for all $m\geq\bar{m}$, we would have

$$\liminf_{M\to\infty}\sum_{m=\bar{m}}^{M}\eta^{(m)}\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2^2\geq\liminf_{M\to\infty}\epsilon^2\sum_{m=\bar{m}}^{M}\eta^{(m)}=\infty$$

which contradicts (54). Therefore, $\liminf\limits_{m\to\infty}\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2=0$. □

8.5.3 Proof of Lemma 4.9

Lemma 8.12 *Let Y_t, W, t and Z_t be three sequences such that W_t is nonnegative for all t. Assume that*

$$Y_{t+1}\leq Y_t-W_t+Z_t,\quad t=0,1,...,$$

and that the series $\sum_{t=0}^{T}Z_t$ converges as $T\to\infty$. Then either $Y_t\to\infty$ or else Y_t converges to a finite value and $\sum_{t=0}^{\infty}W_t<\infty$.

This lemma has been proven by Bertsekas and Tsitsiklis [4].

Lemma 8.13 *When*

$$\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} < \infty,$$

it follows that $\bar{f}(\theta^{(m)}, \bar{\lambda})$ *converge to a finite value.*

Proof. By Proposition 8.8, we have

$$\bar{f}(\theta^{(m+1)}, \bar{\lambda}) \leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - \eta^{(m)}\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2 + \eta^{(m)}\bar{L}M\sqrt{n_2}a_m + \frac{1}{2}(\eta^{(m)})^2 \cdot N\bar{L}M.$$

Let $Y^{(m)} := \bar{f}(\theta^{(m)}, \bar{\lambda})$, $W^{(m)} := \eta^{(m)}\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|_2^2$ and $Z^{(m)} := \eta^{(m)}\bar{L}M\sqrt{n_2}a_m + \frac{1}{2}(\eta^{(m)})^2 \cdot N\bar{L}M$. By (2) and (51)- (53), it is easy to see that $\sum_{m=0}^{M} Z^{(m)}$ converges as $M \to \infty$. Therefore, by Lemma 8.12, $Y^{(m)}$ converges to a finite value. The infinite case can not occur in our setting due to Assumptions 4.1 and 4.2. □

Lemma 8.14 *If*

$$\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} < \infty,$$

then $\lim_{m\to\infty}\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 = 0.$

Proof. To show that $\lim_{m\to\infty}\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 = 0$, assume the contrary; that is,

$$\limsup_{m\to\infty}\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 > 0.$$

Then there exists an $\epsilon > 0$ such that $\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\| < \epsilon/2$ for infinitely many m and also $\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\| > \epsilon$ for infinitely many m. Therefore, there is an infinite subset of integers $\mathbb{M}$, such that for each $m \in \mathbb{M}$, there exists an integer $q(m) > m$ such that

$$\begin{aligned}
\|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\| &< \epsilon/2, \\
\|\nabla\bar{f}(\theta^{(i(m))}, \bar{\lambda})\| &> \epsilon, \\
\epsilon/2 \leq \|\nabla\bar{f}(\theta^{(i)}, \bar{\lambda})\| &\leq \epsilon, \\
\text{if } m < i < q(m).&
\end{aligned} \tag{55}$$

From

$$\begin{aligned}
\|\nabla\bar{f}(\theta^{(m+1)}, \bar{\lambda})\| - \|\nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\| &\leq \|\nabla\bar{f}(\theta^{(m+1)}, \bar{\lambda}) - \nabla\bar{f}(\theta^{(m)}, \bar{\lambda})\| \\
&\leq \bar{L}\|\theta^{(m+1)} - \theta^{(m)}\| \\
&= \bar{L}\eta^{(m)}\|\nabla\bar{f}(\theta^{(m)}, \lambda^{(m)})\|,
\end{aligned}$$

it follows that for all $m \in \mathbb{M}$ that are sufficiently large so that $\bar{L}\eta^{(m)} < \epsilon/4$, we have

$$\epsilon/4 \leq \|\nabla\bar{f}(\theta^{(m)}, \lambda^{(m)})\|. \tag{56}$$

Otherwise the condition $\epsilon/2 \leq \|\nabla\bar{f}(\theta^{(m+1)}, \bar{\lambda})\|$ would be violated. Without loss of generality, we assume that the above relations as well as (37) hold for all $m \in \mathbb{M}$. With the above observations, we have for all $m \in \mathbb{M}$,

$$
\begin{aligned}
\frac{\epsilon}{2} &\leq \|\nabla \bar{f}(\theta^{q(m)}, \bar{\lambda})\| - \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\| \\
&\leq \|\nabla \bar{f}(\theta^{q(m)}, \bar{\lambda}) - \nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\| \\
&\leq \bar{L}\|\theta^{q(m)} - \theta^{(m)}\| \\
&\leq \bar{L} \sum_{i=m}^{q(m)-1} \|\theta^{(i+1)} - \theta^{(i)}\| \\
&\leq \bar{L} \sum_{i=m}^{q(m)-1} \eta^{(i)} \|\nabla \bar{f}(\theta^{(i)}, \lambda^{(i)})\| \\
&\leq \bar{L} \sum_{i=m}^{q(m)-1} \eta^{(i)} (\|\nabla \bar{f}(\theta^{(i)}, \bar{\lambda})\| + \|\nabla \bar{f}(\theta^{(i)}, \lambda^{(i)}) - \nabla \bar{f}(\theta^{(i)}, \bar{\lambda})\|) \\
&\leq \bar{L} \sum_{i=m}^{q(m)-1} \eta^{(i)} (\|\nabla \bar{f}(\theta^{(i)}, \bar{\lambda})\| + \bar{L}\sqrt{n_2} a_m) \\
&\leq \bar{L}\epsilon \sum_{i=m}^{q(m)-1} \eta^{(i)} + \bar{L}^2 \sqrt{n_2} \sum_{i=m}^{q(m)-1} \eta^{(i)} a_m \\
&= \bar{L}\epsilon \sum_{i=m}^{q(m)-1} \eta^{(i)} + \bar{L}^2 \sqrt{n_2} \sum_{i=m}^{q(m)-1} \eta^{(i)} \left(M_1 \sum_{j=m}^{\infty} \sum_{k=1}^{j} \alpha^{(j)} \eta^{(k)} + M_2 \sum_{j=m}^{\infty} \alpha^{(j)} \right) \\
&= \bar{L}\epsilon \sum_{i=m}^{q(m)-1} \eta^{(i)} + \bar{L}^2 \sqrt{n_2} M_1 \sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \sum_{k=1}^{j} \alpha^{(j)} \eta^{(k)} + \bar{L}^2 \sqrt{n_2} M_2 \sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \alpha^{(j)}
\end{aligned}
$$

The first inequality is by (55) and the third one is by the Lipschitz condition assumption. The seventh one is by (33). By (3), we have for all $m \in \mathbb{M}$,

$$\sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \sum_{k=1}^{j} \alpha^{(j)} \eta^{(k)} < \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \sum_{k=1}^{j} \alpha^{(j)} \eta^{(k)} < \infty$$

and

$$\sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \alpha^{(j)} < \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \alpha^{(j)} < \infty.$$

It is easy to see that for any sequence $\{\alpha_i\}$ with $\sum_{i=1}^{\infty} \alpha_i < \infty$, if follows that $\liminf_{M \to \infty} \sum_{i=M}^{\infty} \alpha_i = 0$. Therefore,

$$\liminf_{m \to \infty} \sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \sum_{k=1}^{j} \alpha^{(j)} \eta^{(k)} = 0$$

and

$$\liminf_{m \to \infty} \sum_{i=m}^{q(m)-1} \eta^{(i)} \sum_{j=m}^{\infty} \alpha^{(j)} = 0.$$

From this it follows that

$$\liminf_{m \to \infty} \sum_{i=m}^{q(m)-1} \eta^{(i)} \geq \frac{1}{2\bar{L}}. \tag{57}$$

By the triangle inequality, we have

$$\begin{aligned}
&\|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\| \\
=&\|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}) - \nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}) + \nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\| \\
\geq &\left| \|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})\| - \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda}) - \nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})\| \right|.
\end{aligned}$$

By (33) and (56), if we pick $m \in \mathbb{M}$ such that $L\sqrt{n_2}a_m \leq \frac{\epsilon}{8}$, we have $\|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\| \geq \frac{\epsilon}{8}$. Using (37), we observe that

$$\begin{aligned}
\bar{f}(\theta^{q(m)}, \bar{\lambda}) &\leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - \sum_{i=m}^{q(m)-1} \left(\eta^{(i)} c_1 \|\nabla \bar{f}(\theta^{(i)}, \bar{\lambda})\|_2^2\right) + \frac{1}{2} \cdot N\bar{L}M \sum_{i=m}^{q(m)-1} (\eta^{(i)})^2 \\
&\leq \bar{f}(\theta^{(m)}, \bar{\lambda}) - c_1 \left(\frac{\epsilon}{8}\right)^2 \sum_{i=m}^{q(m)-1} \eta^{(i)} + \frac{1}{2} \cdot N\bar{L}M \sum_{i=m}^{q(m)-1} (\eta^{(i)})^2, \forall m \in \mathbb{M},
\end{aligned}$$

where the second inequality is by (56). By Lemma 8.13, $\bar{f}(\theta^{q(m)}, \bar{\lambda})$ and $\bar{f}(\theta^{(m)}, \bar{\lambda})$ converge to the same finite value. Using this convergence result and the assumption $\sum_{m=0}^{\infty}(\eta^{(m)})^2 < \infty$, this relation implies that

$$\limsup_{m\to\infty, m\in\mathbb{M}} \sum_{i=m}^{q(m)-1} \eta^{(i)} = 0$$

and contradicts (57). □

By Lemmas 8.12, 8.13 and 8.14, we show that Theorem 4.10 holds. To this end we write

$$\begin{aligned}
&\lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})\|_2 \\
\leq &\lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)}) - \nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 + \lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 \\
\leq &\lim_{m\to\infty} \bar{L}\|\lambda^{(m)} - \bar{\lambda}\|_2 + \lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2.
\end{aligned}$$

By Theorem 4.6, we have

$$\lim_{m\to\infty} \bar{L}\|\lambda^{(m)} - \bar{\lambda}\|_2 = 0$$

and by Lemma 4.9, we have

$$\lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \bar{\lambda})\|_2 = 0.$$

Therefore, we have

$$\lim_{m\to\infty} \|\nabla \bar{f}(\theta^{(m)}, \lambda^{(m)})\|_2^2 = 0,$$

which is the statement in Theorem 4.10.

8.6 Proof of Theorem 5.2

In this section we assume that $f_i(\cdot)$ is strongly convex.

Lemma 8.15 *If*

$$m^2 \sum_{i=m}^{\infty} \sum_{n=1}^{i} \alpha^{(i)} \eta^{(n)} < \infty, \tag{58}$$

there exists a constant M such that , for every m we have

$$\frac{|\Delta_1^{(m+1)}| + (1 + \eta^{(m)}c)|\Delta_1^{(m)}| + \eta^{(m)}\bar{L}M\sqrt{n_2}a_m}{\frac{1}{2}{\eta^{(m)}}^2} \leq M. \tag{59}$$

Proof. The notation here is the same as the one used in the proof of Lemma 8.11. Showing (59) is equivalent to showing constant upper bounds for $\frac{|\Delta_1^{(m)}|}{{\eta^{(m)}}^2}$ and $\frac{a_m}{\eta^{(m)}}$.

For an upper bound of $\frac{|\Delta_1^{(m)}|}{{\eta^{(m)}}^2}$, by (41) and (44), we have

$$\begin{aligned}&\frac{|\Delta_1^{(m)}|}{{\eta^{(m)}}^2}\\ =&\frac{|\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{(m)},\bar{\lambda})|}{{\eta^{(m)}}^2}\\ \leq&\frac{M_3}{{\eta^{(m)}}^2}\sum_{j=1}^{D}\left(|k^{(1)}|NDM\frac{1}{\epsilon_B^2}\left|\bar{\sigma}_j-\sigma_j^{(m)}\right|+N\left|\frac{\bar{\mu}_j}{\bar{\sigma}_j+\epsilon_B}-\frac{\mu_j^{(m)}}{\sigma_j^{(m)}+\epsilon_B}\right|\right).\end{aligned} \tag{60}$$

We can see that it is equivalent to show that $\frac{|\bar{\sigma}_j-\sigma_j^{(m)}|}{{\eta^{(m)}}^2}$ and $\frac{|\bar{\mu}_j-\mu_j^{(m)}|}{{\eta^{(m)}}^2}$ have constant upper bounds because all other terms in the right-hand side of (60) are finite constants.

By (39), we have

$$|\bar{\sigma}_j-\sigma_j^{(m)}|\leq\frac{|k^{(1)}|\cdot\tilde{M}_{\bar{L},M}}{C}\sum_{i=m}^{\infty}\alpha^{(i)}\sum_{j=1}^{i}\eta^{(j)}+\frac{\bar{\sigma}_j+|k^{(1)}||\bar{C}|}{C}\sum_{n=m+1}^{\infty}\alpha^{(n)}.$$

Note that we have $\eta^{(m)}=\frac{\zeta}{\vartheta+m}$ and thus ${\eta^{(m)}}^2=O(\frac{1}{m^2})$. Therefore, (58) implies that

$$\frac{1}{{\eta^{(m)}}^2}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)}<\infty. \tag{61}$$

Inequality (61) implies

$$\frac{1}{{\eta^{(m)}}^2}\sum_{i=m+1}^{\infty}\alpha^{(i)}<\infty. \tag{62}$$

This is by the fact that we assume $\sum_{n=1}^{\infty}\eta^{(n)}=\infty$ in Assumption 4.3. We now apply the same kind of analysis to $|\bar{\mu}_j-\mu_j^{(m)}|$ to establish

$$\begin{aligned}&\left|\mu_j^{(m)}-\bar{\mu}_j\right|\\ =&(1-\alpha^{(1)})...(1-\alpha^{(m)})\left|\frac{\mu_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}-\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}\right|\\ \leq&(1-\alpha^{(1)})...(1-\alpha^{(m)})[\left|\frac{\mu_j^{(m)}}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}-\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(\infty)})}\right|+\\ &\left|\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}-\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(\infty)})}\right|]\\ \leq&\left|\tilde{\mu}_j^{(m)}-\tilde{\mu}_j^{(\infty)}\right|+\left|\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}-\tilde{\mu}_j^{(\infty)}\right|.\end{aligned} \tag{63}$$

We define $A_m:=\left|\tilde{\mu}_j^{(m)}-\tilde{\mu}_j^{(\infty)}\right|$ and $B_m:=\left|\frac{\bar{\mu}_j}{(1-\alpha^{(1)})...(1-\alpha^{(m)})}-\tilde{\mu}_j^{(\infty)}\right|$. Recall from Theorem 4.6 that $\{\mu_j^{(m)}\}$ is a Cauchy series. By (8), we have

$$|\tilde{\mu}_j^{(p)} - \tilde{\mu}_j^{(q)}| \le \tilde{M}_{\bar{L},M} \cdot \sum_{m=p}^{q} \sum_{n=1}^{m} \alpha^{(m)} \eta^{(n)}.$$

Therefore, the first term in (63) is bounded by

$$|\tilde{\mu}_j^{(m)} - \tilde{\mu}_j^{\infty}| \le \tilde{M}_{\bar{L},M} \cdot \sum_{i=m}^{\infty} \sum_{n=1}^{i} \alpha^{(i)} \eta^{(n)} < \infty.$$

For the second term in (63), we first define $C := (1 - \alpha^{(1)})...(1 - \alpha^{(u)})...$. Then we have

$$\begin{aligned} & C \cdot \left| \frac{\bar{\mu}_j}{(1 - \alpha^{(1)})...(1 - \alpha^{(m)})} - \tilde{\mu}_j^{(\infty)} \right| \\ =& \bar{\mu}_j |1 - (1 - \alpha^{(m+1)})...(1 - \alpha^{(\infty)})| \\ \le& \bar{\mu}_j \sum_{i=m+1}^{\infty} \alpha^{(i)}, \end{aligned}$$

where the last inequality can be easily checked by induction. Therefore, the second term in (63) is bounded by

$$\left| \frac{\bar{\mu}_j}{(1 - \alpha^{(1)})...(1 - \alpha^{(m)})} - \tilde{\mu}_j^{(\infty)} \right| \le \frac{\bar{\mu}_j}{C} \sum_{i=m+1}^{\infty} \alpha^{(i)}.$$

Hence (61) and (62) ensure $\dfrac{|\bar{\mu}_j - \mu_j^{(m)}|}{{\eta^{(m)}}^2}$ to be finite.

For an upper bound of $\dfrac{a_m}{\eta^{(m)}}$, by (24), we have

$$\frac{a_m}{\eta^{(m)}} = \frac{M_1 \sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \eta^{(j)} + M_2 \sum_{i=m}^{\infty} \alpha^{(i)}}{\eta^{(m)}}.$$

We know that

$$\frac{M_1 \sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \eta^{(j)}}{\eta^{(m)}} < M_1 \frac{1}{{\eta^{(m)}}^2} \sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \eta^{(j)} < \infty \tag{64}$$

and

$$\frac{M_2 \sum_{i=m}^{\infty} \alpha^{(i)}}{\eta^{(m)}} < M_2 \frac{1}{{\eta^{(m)}}^2} \sum_{i=m}^{\infty} \alpha^{(i)} < \infty. \tag{65}$$

The second inequalities in (64) and (65) are by (61) and (62). Note that given that $\eta^{(m)} = 1/m$, (61) is equivalent to

$$\sum_{i=m}^{\infty} \sum_{j=1}^{i} \alpha^{(i)} \frac{1}{j} < \sum_{i=m}^{\infty} \alpha^{(i)} \ln(i) < \sum_{i=1}^{\infty} \alpha^{(i)} \ln(i) < \infty$$

This concludes the proof. □

Lemma 8.16 *Under the assumptions of Lemma 8.15, Theorem 5.2 holds.*

The proof for this Lemma of the high level follows the proof of Theorem 4.7 in Bottou et al. [5].

Proof. Assumption 5.1 implies that

$$\bar{f}(\tilde{\theta}, \lambda) \ge \bar{f}(\hat{\theta}, \lambda) + \nabla \bar{f}(\hat{\theta}, \lambda)^T (\tilde{\theta} - \hat{\theta}) + \frac{1}{2} c \|\tilde{\theta} - \hat{\theta}\|_2^2, \forall \tilde{\theta}, \hat{\theta}.$$

Therefore, $\bar{f}$ has a unique minimizer $\bar{f}^* := \bar{f}(\theta^*, \lambda)$ for any λ fixed. Note that $\theta^* = \theta^*(\lambda)$ but this dependency is irrelevant in the rest of the proof. Standard convex analysis argument establishes

$$2c\left(\bar{f}(\theta^{(m)},\lambda)-\bar{f}(\theta^{*},\lambda)\right)\leq\|\nabla\bar{f}(\theta^{(m)},\lambda)\|_2^2. \tag{66}$$

Recall that $\Delta_1^{(m+1)}:=\bar{f}(\theta^{(m+1)},\lambda^{(m+1)})-\bar{f}(\theta^{(m+1)},\bar{\lambda})$. We then have

$$\begin{aligned}
&\bar{f}(\theta^{(m+1)},\lambda^{(m+1)})-\bar{f}(\theta^{(m)},\lambda^{(m)})\\
=&\left[\bar{f}(\theta^{(m+1)},\lambda^{(m+1)})-\bar{f}(\theta^{(m+1)},\bar{\lambda})\right]-\left[\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{(m)},\bar{\lambda})\right]\\
&+\bar{f}(\theta^{(m+1)},\bar{\lambda})-\bar{f}(\theta^{(m)},\bar{\lambda})\\
\leq&|\Delta_1^{(m+1)}|+|\Delta_1^{(m)}|+\bar{f}(\theta^{(m+1)},\bar{\lambda})-\bar{f}(\theta^{(m)},\bar{\lambda}).
\end{aligned} \tag{67}$$

Therefore,

$$\begin{aligned}
&\bar{f}(\theta^{(m+1)},\bar{\lambda})-\bar{f}(\theta^{(m)},\bar{\lambda})\\
\leq&-\eta^{(m)}\|\nabla\bar{f}(\theta^{(m)},\bar{\lambda})\|_2^2+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M\\
\leq&-\eta^{(m)}c(\bar{f}(\theta^{(m)},\bar{\lambda})-\bar{f}(\theta^{*},\bar{\lambda}))+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M\\
=&-\eta^{(m)}c\left(\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{*},\bar{\lambda})+\bar{f}(\theta^{(m)},\bar{\lambda})-\bar{f}(\theta^{(m)},\lambda^{(m)})\right)\\
&+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M\\
\leq&-\eta^{(m)}c\left(\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{*},\bar{\lambda})\right)+\eta^{(m)}c\left|\bar{f}(\theta^{(m)},\bar{\lambda})-\bar{f}(\theta^{(m)},\lambda^{(m)})\right|\\
&+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M\\
=&-\eta^{(m)}c\left(\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{*},\bar{\lambda})\right)+\eta^{(m)}c|\Delta_1^{(m)}|\\
&+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M.
\end{aligned} \tag{68}$$

The first inequality is by Proposition 8.8, while the second inequality is by the strong convexity property (66). Combining (67) and (68) yields

$$\begin{aligned}
&\bar{f}(\theta^{(m+1)},\lambda^{(m+1)})-\bar{f}(\theta^{(m)},\lambda^{(m)})\\
\leq&-\eta^{(m)}c\left(\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{*},\bar{\lambda})\right)+|\Delta_1^{(m+1)}|+(1+\eta^{(m)}c)|\Delta_1^{(m)}|\\
&+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m+\frac{1}{2}{\eta^{(m)}}^2N\bar{L}M.
\end{aligned}$$

By Lemma 8.15, there exists an upper bound M_4 such that for all m sufficiently large,

$$\frac{|\Delta_1^{(m+1)}|+(1+\eta^{(m)}c)|\Delta_1^{(m)}|+\eta^{(m)}\bar{L}M\sqrt{n_2}a_m}{\frac{1}{2}{\eta^{(m)}}^2}\leq M_4.$$

By subtracting $\bar{f}(\theta^{*},\bar{\lambda})$ from both side of (8.6), we obtain

$$\begin{aligned}
&\bar{f}(\theta^{(m+1)},\lambda^{(m+1)})-\bar{f}(\theta^{*},\bar{\lambda})\\
\leq&(1-\eta^{(m)}c)(\bar{f}(\theta^{(m)},\lambda^{(m)})-\bar{f}(\theta^{*},\bar{\lambda}))+\frac{1}{2}{\eta^{(m)}}^2(N\bar{L}M+M_4).
\end{aligned} \tag{69}$$

Inequality (69) has the exact same form used in classic convergence proofs for the strongly convex, diminishing step size case.

We finally show by induction that

$$\bar{f}(\theta_m,\lambda_m)-\bar{f}(\theta^{*},\bar{\lambda})\leq\frac{v}{\vartheta+m} \tag{70}$$

holds for all m, where

$$v := \max\{\frac{\zeta^2(N\bar{L}M + M_4)}{2(\zeta c - 1)}, (\vartheta + 1)[\bar{f}(\theta^{(1)}, \lambda^{(1)}) - \bar{f}(\theta^*, \bar{\lambda})]\}.$$

First, the definition of ζ ensures that it holds for $m = 1$. Assuming (70) holds for some $m \geq 1$, it follows from (69) that

$$\begin{aligned}
\bar{f}(\theta^{(m+1)}, \lambda^{(m+1)}) - \bar{f}(\theta^*, \bar{\lambda}) &\leq (1 - \eta^{(m)}c)(\bar{f}(\theta^{(m)}, \lambda^{(m)}) - \bar{f}(\theta^*, \bar{\lambda})) + \frac{1}{2}{\eta^{(m)}}^2(LM + M_4) \\
&\leq (1 - \eta^{(m)}c)\frac{v}{\vartheta + m} + \frac{1}{2}{\eta^{(m)}}^2(LM + M_4) \\
&= (1 - \frac{\zeta c}{\vartheta + m})\frac{v}{\vartheta + m} + \frac{\zeta^2(LM + M_4)}{2(\vartheta + m)^2} \\
&= \frac{\vartheta + m - \zeta c}{(\vartheta + m)^2}v + \frac{\zeta^2(LM + M_4)}{2(\vartheta + m)^2} \\
&= \frac{\vartheta + m - 1}{(\vartheta + m)^2}v - \left(\frac{\zeta c - 1}{(\vartheta + m)^2}v\right) + \frac{\zeta^2(LM + M_4)}{2(\vartheta + m)^2} \\
&\leq \frac{\vartheta + m - 1}{(\vartheta + m)^2}v \\
&\leq \frac{v}{\vartheta + m + 1}.
\end{aligned}$$

The first inequality is by (69), the second inequality is by the definition of $\eta^{(m)}$, the third inequality is by the definition of v, the sum of the latter two terms is non-positive, and the fourth inequality is because $(\vartheta + m)^2 \geq (\vartheta + m + 1)(\vartheta + m - 1)$. This shows that the algorithm converges at a sublinear rate. $\square$

9 Appendix B

9.1 Conditions for stepsizes

Here we discuss the actual conditions for $\eta^{(m)}$ and $\alpha^{(m)}$ to satisfy the assumptions of Theorem 4.6, Lemma 4.7 and Theorem 5.2, respectively. We only consider the cases $\eta^{(m)} = \frac{1}{m^k}$ and $\alpha^{(m)} = \frac{1}{m^h}$, but the same analysis applies to the cases $\eta^{(m)} = O(\frac{1}{m^k})$ and $\alpha^{(m)} = O(\frac{1}{m^h})$.

9.1.1 Assumptions of Theorem 4.6

For the assumptions of Theorem 4.6, the first condition

$$\sum_{m=1}^{\infty} \alpha^{(m)} < \infty$$

requires $h > 1$.

Besides, the second condition

$$\begin{aligned}
\sum_{m=1}^{\infty}\sum_{n=1}^{m} \alpha^{(m)}\eta^{(n)} = \sum_{n=1}^{\infty}\sum_{m=n}^{\infty} \alpha^{(m)}\eta^{(n)} = \sum_{n=1}^{\infty} \eta^{(n)} \sum_{m=n}^{\infty} \alpha^{(m)} \\
\approx \frac{1}{h-1}\sum_{n=1}^{\infty} \eta^{(n)} \frac{1}{n^{h-1}} = \frac{1}{h-1}\sum_{n=1}^{\infty} \frac{1}{n^{k+h-1}} < \infty
\end{aligned}$$

requires $k + h > 2$. The approximation comes from the fact that for every $p > 1$, we have

$$\sum_{k=n}^{\infty} k^{-p} \approx \int_{k=n}^{\infty} k^{-p} dx = \frac{1}{1-p} x^{1-p}\Big|_n^{\infty} = \frac{1}{p-1}\frac{1}{n^{p-1}}.$$

Since $k \geq 1$ due to Assumption 4.3, we conclude that $k + h > 2$.

Therefore, the conditions for $\eta^{(m)}$ and $\alpha^{(m)}$ to satisfy the assumptions of Theorem 4.6 are $h > 1$ and $k \geq 1$.

9.1.2 Assumptions of Lemma 4.7

For the assumptions of Theorem 4.6, the first condition

$$\sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\alpha^{(n)} \approx \sum_{m=1}^{\infty}\frac{1}{m^{h-1}} < \infty$$

requires $h > 2$.

Besides, the second condition is

$$\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\sum_{n=1}^{i}\alpha^{(i)}\eta^{(n)} = \sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\alpha^{(i)}\sum_{n=1}^{i}\eta^{(n)} \leq C\sum_{m=1}^{\infty}\sum_{i=m}^{\infty}\alpha^{(i)} < \infty.$$

The inequality holds because for any $p > 1$, we have

$$\sum_{k=1}^{n}k^{-p} \approx \int_{k=1}^{n}k^{-p}dk = \frac{1}{1-p}k^{1-p}\bigg|_{1}^{n} = \frac{1}{p-1}(1-n^{1-p}) \leq C$$

Therefore, the conditions for $\eta^{(m)}$ and $\alpha^{(m)}$ to satisfy the assumptions of Lemma 4.7 are $h > 2$ and $k \geq 1$.

9.1.3 Assumptions of Theorem 5.2

Recall that we have let $\eta^{(m)} = 1/m$. For the assumptions of Theorem 5.2, the condition

$$\sum_{m=1}^{\infty}\alpha^{(m)}\ln(m) < \infty$$

requires $h > 1$. To see this, note that $\ln(m) \leq Cm^{\epsilon}$ for any $\epsilon > 0$. Thus

$$\sum_{m=1}^{\infty}\alpha^{(m)}\ln(m) \leq C\sum_{m=1}^{\infty}m^{-h}m^{\epsilon} < \infty$$

if $\epsilon - h < -1$. This yields $h > 1$.

Therefore, the condition for $\alpha^{(m)}$ to satisfy the assumptions of Theorem 5.2 is $h > 1$.