Mitigating Gradient Staleness in Decoupled Learning for Classification Tasks

Huiping Zhuang, Zhenyu Weng, Fulin Luo, Kar-Ann Toh, Haizhou Li, Zhiping Lin*

SUPPLEMENTARY MATERIAL A: PROOF OF THEOREM 1

Proof. To simplify the notations, let $\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} = \frac{1}{M}\sum_{j=0}^{M-1} \boldsymbol{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s+j-2(K-k)}$ and $\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} = \frac{1}{M}\sum_{j=0}^{M-1} \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s+j-2(K-k)}$. According to Assumption 1, the following inequality holds:

$$
\begin{aligned}
f(\boldsymbol{\theta}^{U_{s+1}}) \leq & f(\boldsymbol{\theta}^{U_s}) + (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s})^T(\boldsymbol{\theta}^{U_{s+1}} - \boldsymbol{\theta}^{U_s}) + \frac{L}{2}\left\|\boldsymbol{\theta}^{U_{s+1}} - \boldsymbol{\theta}^{U_s}\right\|_2^2 \\
= & f(\boldsymbol{\theta}^{U_s}) - \gamma_s \sum_{k=1}^{K} (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T \mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} + \frac{L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}\right\|_2^2
\end{aligned} \tag{1}
$$

which can be further developed such that

$$
\begin{aligned}
f(\boldsymbol{\theta}^{U_{s+1}}) \leq & f(\boldsymbol{\theta}^{U_s}) - \gamma_s \sum_{k=1}^{K} (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T (\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s} + \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}) + \frac{L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s} + \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
= & f(\boldsymbol{\theta}^{U_s}) - \gamma_s \sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 - \gamma_s \sum_{k=1}^{K} (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T (\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}) + \frac{L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
& + \frac{L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 + L\gamma_s^2 \sum_{k=1}^{K} (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T (\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}) \\
= & f(\boldsymbol{\theta}^{U_s}) - \left(\gamma_s - \frac{L\gamma_s^2}{2}\right)\sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 + \tilde{Q}_1 + \tilde{Q}_2
\end{aligned} \tag{2}
$$

where

$$
\tilde{Q}_1 = \frac{L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2,\ \tilde{Q}_2 = (L\gamma_s^2 - \gamma_s)\sum_{k=1}^{K} (\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T (\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}).
$$

The expectation of $\tilde{Q}_1$ is bounded by

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\{\tilde{Q}_1\} = & \frac{L\gamma_s^2}{2}\mathbb{E}_{\boldsymbol{x}}\left\{\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2\right\} = \frac{L\gamma_s^2}{2}\mathbb{E}_{\boldsymbol{x}}\left\{\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s} + \bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}\right\|_2^2\right\} \\
\leq & L\gamma_s^2\mathbb{E}_{\boldsymbol{x}}\left\{\sum_{k=1}^{K}\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}\right\|_2^2\right\} + L\gamma_s^2\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
= & L\gamma_s^2\mathbb{E}_{\boldsymbol{x}}\left\{\left\|\mathfrak{g}_{\boldsymbol{\theta}}^{U_s^{'}} - \bar{\mathfrak{g}}_{\boldsymbol{\theta}}^{U_s^{'}}\right\|_2^2\right\} + L\gamma_s^2\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
\leq & L\gamma_s^2\mathbb{E}_{\boldsymbol{x}}\left\{\left\|\mathfrak{g}_{\boldsymbol{\theta}}^{U_s^{'}}\right\|_2^2\right\} + L\gamma_s^2\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
\leq & L\gamma_s^2\frac{1}{M^2}\mathbb{E}_{\boldsymbol{x}}\left\{\sum_{j=0}^{M-1}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s+j-2(K-k)}\right\|_2^2\right\} + L\gamma_s^2\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 \\
\leq & \frac{AL}{M}\gamma_s^2 + L\gamma_s^2\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}} - \bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2 = \frac{AL}{M}\gamma_s^2 + L\gamma_s^2\tilde{P}_1
\end{aligned}
$$

Corresponding author: Zhiping Lin. *Email address*: EZPLin@ntu.edu.sg

where the first inequality follows from $\|\boldsymbol{x}+\boldsymbol{y}\|_2^2\leq 2\|\boldsymbol{x}\|_2^2+2\|\boldsymbol{y}\|_2^2$. The second inequality is from $\mathbb{E}\{\|\epsilon-\mathbb{E}\{\epsilon\}\|_2^2\}\leq \mathbb{E}\{\|\epsilon\|_2^2\}-\|\mathbb{E}\{\epsilon\}\|_2^2\leq \mathbb{E}\{\|\epsilon\|_2^2\}$ due to gradient unbiasedness (i.e., $\mathbb{E}_{\boldsymbol{x}}\{\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}\}=\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}$). The last inequality follows from Assumption 2, and $\tilde{P}_1$ is bounded by

$$
\begin{aligned}
\tilde{P}_1=&\sum_{k=1}^{K}\left\|\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}-\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2\leq\frac{1}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s+j-2(K-k)}-\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2\\
=&\frac{L^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\left\|\boldsymbol{\theta}_{q(k)}^{U_s}-\boldsymbol{\theta}_{q(k)}^{U_{s-d_{k,j}}}\right\|_2^2\\
=&\frac{L^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\left\|\sum_{\alpha=\max\{0,s-d_{k,j}\}}^{s-1}(\boldsymbol{\theta}_{q(k)}^{U_{\alpha+1}}-\boldsymbol{\theta}_{q(k)}^{U_{\alpha}})\right\|_2^2\leq\frac{L^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\sum_{\alpha=\max\{0,s-d_{k,j}\}}^{s-1}\left\|\boldsymbol{\theta}_{q(k)}^{U_{\alpha+1}}-\boldsymbol{\theta}_{q(k)}^{U_{\alpha}}\right\|_2^2\\
=&\frac{L^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\sum_{\alpha=\max\{0,s-d_{k,j}\}}^{s-1}\gamma_\alpha^2\left\|\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}\right\|_2^2\leq\frac{L^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\sum_{\alpha=\max\{0,s-d_{k,j}\}}^{s-1}\gamma_\alpha^2\frac{1}{M^2}\sum_{j=0}^{M-1}\left\|\boldsymbol{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s+j-2(K-k)}\right\|_2^2\\
\leq&\frac{AL^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}\sum_{\alpha=\max\{0,s-d_{k,j}\}}^{s-1}\gamma_\alpha^2\leq\gamma_s^2\frac{AL^2}{M^2}\sum_{k=1}^{K}\sum_{j=0}^{M-1}(s-\max\{0,s-d_{k,j}\})\\
\leq&\gamma_s^2\frac{AL^2}{M^3}\sum_{k=1}^{K}\sum_{j=0}^{M-1}d_{k,j}=\gamma_s^2\frac{AL^2}{M^2}\sum_{k=1}^{K}\bar{d}_k
\end{aligned}
$$

with the first inequality coming from Assumption 1. On the other hand, the expectation of $\tilde{Q}_2$ is bounded by

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\{\tilde{Q}_2\}=&-(\gamma_s-L\gamma_s^2)\mathbb{E}_{\boldsymbol{x}}\left\{\sum_{k=1}^{K}(\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T\left(\mathfrak{g}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}-\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right)\right\}=-(\gamma_s-L\gamma_s^2)\sum_{k=1}^{K}(\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s})^T\left(\bar{\mathfrak{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s^{'}}-\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right)\\
\leq&\frac{\gamma_s-L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2+\frac{\gamma_s-L\gamma_s^2}{2}\tilde{P}_1
\end{aligned}
$$

where the second equality follows by the unbiased gradient using SGD, and the inequality comes from $\pm\boldsymbol{x}^T\boldsymbol{y}\leq\frac{1}{2}\|\boldsymbol{x}\|_2^2+\frac{1}{2}\|\boldsymbol{y}\|_2^2$.

Taking the expectation of both sides in Eq. (2) and substituting $\tilde{Q}_1$ and $\tilde{Q}_2$, the inequality is rewritten as

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{\theta}^{U_{s+1}})\}\leq&f(\boldsymbol{\theta}^{U_s})-\left(\gamma_s-\frac{L\gamma_s^2}{2}\right)\sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2+\frac{AL}{M}\gamma_s^2+L\gamma_s^2\tilde{P}_1+\frac{\gamma_s-L\gamma_s^2}{2}\sum_{k=1}^{K}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}_{q(k)}}^{U_s}\right\|_2^2+\frac{\gamma_s-L\gamma_s^2}{2}\tilde{P}_1\\
=&f(\boldsymbol{\theta}^{U_s})-\frac{\gamma_s}{2}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2+\frac{\gamma_s+L\gamma_s^2}{2}\tilde{P}_1+\frac{AL}{M}\gamma_s^2\\
\leq&f(\boldsymbol{\theta}^{U_s})-\frac{\gamma_s}{2}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2+\frac{\gamma_s+L\gamma_s^2}{2}\gamma_s^2\frac{AL^2}{M^2}\sum_{k=1}^{K}\bar{d}_k+\frac{AL}{M}\gamma_s^2\\
=&f(\boldsymbol{\theta}^{U_s})-\frac{\gamma_s}{2}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2+\gamma_s^2\Big(\frac{AL}{M}+\frac{\gamma_s+L\gamma_s^2}{2}L\frac{AL}{M^2}\sum_{k=1}^{K}\bar{d}_k\Big)\\
\leq&f(\boldsymbol{\theta}^{U_s})-\frac{\gamma_s}{2}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2+\gamma_s^2\frac{AL}{M}\Big(1+\frac{1}{M}\sum_{k=1}^{K}\bar{d}_k\Big)
\end{aligned}
\tag{3}
$$

where the last inequality follows from $L\gamma_s\leq 1$ such that $\frac{\gamma_s+L\gamma_s^2}{2}L=\frac{1}{2}(L\gamma_s+(L\gamma_s)^2)\leq 1$. The proof is now completed. □

Supplementary material B: Proof of Theorem 2

Proof. By moving $\frac{\gamma_s}{2}\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2$ and $\mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{\theta}^{U_{s+1}})\}$ to the LHS and the RHS of Eq. (24) respectively, and multiplying both sides by 2, we have

$$
\gamma_s\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\leq 2(f(\boldsymbol{\theta}^{U_s})-\mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{\theta}^{U_{s+1}})\})+2\gamma_s^2\frac{AL}{M}\Big(1+\frac{1}{M}\sum_{k=1}^{K}\bar{d}_k\Big).
\tag{4}
$$

Take full expectation on both sides of Eq. (4), and it leads to

$$\gamma_s \mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\} \leq 2(\mathbb{E}\{f(\boldsymbol{\theta}^{U_s})\} - \mathbb{E}\{f(\boldsymbol{\theta}^{U_{s+1}})\}) + 2\gamma_s^2 \frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big). \tag{5}$$

By summing both sides of Eq. (5) from 0 to $S-1$, and dividing it by $\mathbb{T}_S = \sum_{s=0}^{S-1} \gamma_s$, it becomes

$$\begin{aligned}\frac{1}{\mathbb{T}_S}\sum_{s=0}^{S-1} \gamma_s \mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\} &\leq \frac{2(f(\boldsymbol{\theta}^0) - \mathbb{E}\{f(\boldsymbol{\theta}^{U_S})\})}{\mathbb{T}_S} + \frac{2\frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big)\sum_{s=0}^{S-1}\gamma_s^2}{\mathbb{T}_S} \\ &\leq \frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{\mathbb{T}_S} + \frac{2\frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big)\sum_{s=0}^{S-1}\gamma_s^2}{\mathbb{T}_S}.\end{aligned}$$

where the last inequality comes from $f(\boldsymbol{\theta}^*) \leq \mathbb{E}\{f(\boldsymbol{\theta}^{U_S})\}$. □

SUPPLEMENTARY MATERIAL C: PROOF OF THEOREM 3

Proof. We start the proof from Eq. (5) as the constant learning rate is a special case in Theorem 2. By setting $\gamma_s = \gamma$, Eq. Eq. (5) is rewritten as

$$\mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\} \leq \frac{2(\mathbb{E}\{f(\boldsymbol{\theta}^{U_s})\} - \mathbb{E}\{f(\boldsymbol{\theta}^{U_{s+1}})\})}{\gamma} + 2\gamma\frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big). \tag{6}$$

Summing both sides of Eq. (6) from $s = 0$ to $S-1$ and dividing them by S, it leads to

$$\begin{aligned}\frac{1}{S}\sum_{s=0}^{S-1} \mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\} &\leq \frac{2(f(\boldsymbol{\theta}^0) - \mathbb{E}\{f(\boldsymbol{\theta}^{U_S})\})}{\gamma S} + 2\gamma\frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big) \\ &\leq \frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{\gamma S} + 2\gamma\frac{AL}{M}\Big(1 + \frac{1}{M}\sum_{k=1}^{K} \bar{d}_k\Big).\end{aligned} \tag{7}$$

Substituting $\gamma = \epsilon\sqrt{M(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))/\Big(SAL(1 + (1/M)\sum_{k=1}^{K}\bar{d}_k)\Big)}$ into Eq. (7), the RHS becomes

$$\begin{aligned}\frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*)) + 2\gamma^2 SAL\Big(1 + (1/M)\sum_{k=1}^{K}\bar{d}_k\Big)/M}{\gamma S} &= \frac{(2 + 2\epsilon^2)(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{S\epsilon\sqrt{M(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))/\Big(SAL(1 + (1/M)\sum_{k=1}^{K}\bar{d}_k)\Big)}} \\ &= \frac{(2 + 2\epsilon^2)}{\epsilon}\sqrt{AL(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))\Big(1 + (1/M)\sum_{k=1}^{K}\bar{d}_k\Big)/(MS)}.\end{aligned}$$

Since the LHS of Eq. (7) is the average of $\mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\}$ for $s = 0, 1, \ldots, S-1$, we have

$$\min_{t \in \{0,1,\ldots,S-1\}} \mathbb{E}\{\left\|\bar{\boldsymbol{g}}_{\boldsymbol{\theta}}^{U_s}\right\|_2^2\} \leq \frac{(2 + 2\epsilon^2)}{\epsilon}\sqrt{AL(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))\Big(1 + (1/M)\sum_{k=1}^{K}\bar{d}_k\Big)/(MS)}$$

which completes the proof. □