

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 $\mathbf{g}_{\theta_{q(k)}}^{U_s'} = \frac{1}{M} \sum_{j=0}^{M-1} \mathbf{g}_{\theta_{q(k)}}^{U_s+j-2(K-k)}$ and $\bar{\mathbf{g}}_{\theta_{q(k)}}^{U_s'} = \frac{1}{M} \sum_{j=0}^{M-1} \bar{\mathbf{g}}_{\theta_{q(k)}}^{U_s+j-2(K-k)}$. According to Assumption 1, the following inequality holds:

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

which can be further developed such that

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

where

$$\tilde{Q}_1 = \frac{L\gamma_s^2}{2} \sum_{k=1}^K \|\mathbf{g}_{\theta_{q(k)}}^{U_s'} - \bar{\mathbf{g}}_{\theta_{q(k)}}^{U_s}\|_2^2, \quad \tilde{Q}_2 = (L\gamma_s^2 - \gamma_s) \sum_{k=1}^K (\bar{\mathbf{g}}_{\theta_{q(k)}}^{U_s})^T (\mathbf{g}_{\theta_{q(k)}}^{U_s'} - \bar{\mathbf{g}}_{\theta_{q(k)}}^{U_s}).$$

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

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

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

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

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

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

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

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

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

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. $\square$

SUPPLEMENTARY MATERIAL B: PROOF OF THEOREM 2

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

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

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

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

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

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

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

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. (5) is rewritten as

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

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

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

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

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

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

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

which completes the proof. $\square$