Supplementary Materials for Paper "Decoupled Parallel Backpropagation with Convergence Guarantee"

Zhouyuan Huo Bin Gu Qian Yang Heng Huang [1]

A. Proof to Lemma 1

Proof: Because the gradient of $f(w)$ is Lipschitz continuous in Assumption 1, the following inequality holds that:

$$f(w^{t+1}) \quad \leq \quad f(w^t) + \nabla f(w^t)^T \left(w^{t+1} - w^t\right) + \frac{L}{2} \left\|w^{t+1} - w^t\right\|_2^2. \tag{1}$$

From the update rule in Algorithm 1, we take expectation on both sides and obtain:

$$\begin{aligned}
\mathbb{E}\left[f(w^{t+1})\right] \quad \leq \quad & f(w^t) - \gamma_t \mathbb{E}\left[\nabla f(w^t)^T \left(\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}}\left(w^{t-K+k}\right)\right)\right] \\
& + \frac{L\gamma_t^2}{2} \mathbb{E}\left\|\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}}(w^{t-K+k})\right\|_2^2 \\
\leq \quad & f(w^t) - \gamma_t \sum_{k=1}^{K} \nabla f(w^t)^T \left(\nabla f_{\mathcal{G}(k)}\left(w^{t-K+k}\right) + \nabla f_{\mathcal{G}(k)}\left(w^t\right) - \nabla f_{\mathcal{G}(k)}\left(w^t\right)\right) \\
& + \frac{L\gamma_t^2}{2} \mathbb{E}\left\|\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}}(w^{t-K+k}) - \nabla f(w^t) + \nabla f(w^t)\right\|_2^2 \\
= \quad & f(w^t) - \gamma_t \left\|\nabla f(w^t)\right\|_2^2 - \gamma_t \sum_{k=1}^{K} \nabla f(w^t)^T \left(\nabla f_{\mathcal{G}(k)}\left(w^{t-K+k}\right) - \nabla f_{\mathcal{G}(k)}\left(w^t\right)\right) \\
& + \frac{L\gamma_t^2}{2} \left\|\nabla f(w^t)\right\|_2^2 + \frac{L\gamma_t^2}{2} \mathbb{E}\left\|\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}}(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2 \\
& + L\gamma_t^2 \sum_{k=1}^{K} \nabla f(w^t)^T \left(\nabla f_{\mathcal{G}(k)}\left(w^{t-K+k}\right) - \nabla f_{\mathcal{G}(k)}\left(w^t\right)\right) \\
= \quad & f(w^t) - \left(\gamma_t - \frac{L\gamma_t^2}{2}\right) \left\|\nabla f(w^t)\right\|_2^2 + \underbrace{\frac{L\gamma_t^2}{2} \mathbb{E}\left\|\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}}(w^{t-K+k}) - \nabla f(w^t)\right\|^2}_{Q_1} \\
& \underbrace{-(\gamma_t - L\gamma_t^2) \sum_{k=1}^{K} \nabla f(w^t)^T \left(\nabla f_{\mathcal{G}(k)}\left(w^{t-K+k}\right) - \nabla f_{\mathcal{G}(k)}\left(w^t\right)\right)}_{Q_2},
\end{aligned} \tag{2}$$

[1]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, United States. Correspondence to: Heng Huang <heng.huang@pitt.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

where the second inequality follows from the unbiased gradient $\mathbb{E}\left[\nabla f_{x_i}(w)\right] = \nabla f(w)$. Because of $\|x+y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$ and $xy \leq \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2$, we have the upper bound of Q_1 and Q_2 as follows:

$$\begin{aligned} Q_1 &= \frac{L\gamma_t^2}{2}\mathbb{E}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k}) - \nabla f(w^t) - \sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) + \sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k})\right\|_2^2 \\ &\leq \underbrace{L\gamma_t^2\,\mathbb{E}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k}) - \sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k})\right\|_2^2}_{Q_3} + \underbrace{L\gamma_t^2\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2}_{Q_4} \end{aligned} \tag{3}$$

$$\begin{aligned} Q_2 &= -(\gamma_t - L\gamma_t^2)\sum_{k=1}^{K}\nabla f(w^t)^T\left(\nabla f_{\mathcal{G}(k)}\left(w^{t-K+k}\right) - \nabla f_{\mathcal{G}(k)}\left(w^t\right)\right) \\ &\leq \frac{\gamma_t - L\gamma_t^2}{2}\left\|\nabla f(w^t)\right\|_2^2 + \frac{\gamma_t - L\gamma_t^2}{2}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2. \end{aligned} \tag{4}$$

As per the equation regarding variance $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|_2^2 = \mathbb{E}\|\xi\|_2^2 - \|\mathbb{E}[\xi]\|_2^2$, we can bound Q_3 as follows:

$$\begin{aligned} Q_3 &= \mathbb{E}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k}) - \sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k})\right\|_2^2 \\ &= \sum_{k=1}^{K}\mathbb{E}\left\|\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k}) - \nabla f_{\mathcal{G}(k)}(w^{t-K+k})\right\|_2^2 \\ &\leq \sum_{k=1}^{K}\mathbb{E}\left\|\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k})\right\|_2^2 \\ &\leq KM, \end{aligned} \tag{5}$$

where the equality follows from the definition of $\nabla f_{\mathcal{G}(k)}(w)$ such that $[\nabla f_{\mathcal{G}(k)}(w)]_j = 0,\ \forall j \notin \mathcal{G}(k)$ and the last inequality is from Assumption 2. We can also get the upper bound of Q_4:

$$\begin{aligned} Q_4 &= \left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2 \\ &= \sum_{k=1}^{K}\left\|\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) - \nabla f_{\mathcal{G}(k)}(w^t)\right\|_2^2 \\ &\leq \sum_{k=1}^{K}\left\|\nabla f(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2 \\ &\leq L^2\sum_{k=1}^{K}\left\|\sum_{j=\max\{0,t-K+k\}}^{t-1}\left(w^{j+1} - w^j\right)\right\|^2 \\ &\leq L^2\gamma_{\max\{0,t-K+1\}}^2 K\sum_{k=1}^{K}\sum_{j=\max\{0,t-K+k\}}^{t-1}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k),x_{(j)}}\left(w^{j-K+k}\right)\right\|_2^2 \\ &\leq KL\gamma_t\frac{\gamma_{\max\{0,t-K+1\}}}{\gamma_t}\sum_{k=1}^{K}\sum_{j=\max\{0,t-K+k\}}^{t-1}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k),x_{(j)}}\left(w^{j-K+k}\right)\right\|_2^2 \\ &\leq L\gamma_t\sigma K^4 M, \end{aligned} \tag{6}$$

where the second inequality is from Assumption 1, the fourth inequality follows from that $L\gamma_t \leq 1$ and the last inequality follows from $\|z_1 + ... + z_r\|_2^2 \leq r(\|z_1\|_2^2 + ... + \|z_r\|_2^2)$, Assumption 2 and $\sigma := \max_t \frac{\gamma_{\max\{0,t-K+1\}}}{\gamma_t}$. Integrating the upper bound of Q_1, Q_2, Q_3 and Q_4 in (2), we have:

$$\begin{aligned}\mathbb{E}\left[f(w^{t+1})\right] - f(w^t) &\leq -\frac{\gamma_t}{2}\left\|\nabla f(w^t)\right\|_2^2 + \gamma_t^2 L \sum_{k=1}^{K} \mathbb{E}\left\|\nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(w^{t-K+k})\right\|^2 \\ &\quad + \frac{\gamma_t + L\gamma_t^2}{2}\left\|\sum_{k=1}^{K}\nabla f_{\mathcal{G}(k)}(w^{t-K+k}) - \nabla f(w^t)\right\|_2^2. \\ &\leq -\frac{\gamma_t}{2}\left\|\nabla f(w^t)\right\|^2 + \gamma_t^2 L M_K, \end{aligned} \tag{7}$$

where we let $M_K = KM + \sigma K^4 M$.

B. Proof to Theorem 1

Proof: When γ_t is constant and $\gamma_t = \gamma$, taking total expectation of (12) in Lemma 1, we obtain:

$$\mathbb{E}\left[f(w^{t+1})\right] - \mathbb{E}\left[f(w^t)\right] \leq -\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 + \gamma^2 L M_K, \tag{8}$$

where $\sigma = 1$ and $M_K = KM + K^4 M$. Summing (8) from $t = 0$ to $T-1$, we have:

$$\mathbb{E}\left[f(w^T)\right] - f(w^0) \leq -\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 + T\gamma^2 L M_K. \tag{9}$$

Suppose w^* is the optimal solution for $f(w)$, therefore $f(w^*) - f(w^0) \leq \mathbb{E}\left[f(w^T)\right] - f(w^0)$. Above all, the following inequality is guaranteed that:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 \leq \frac{2\left(f(w^0) - f(w^*)\right)}{\gamma T} + 2\gamma L M_K. \tag{10}$$

C. Proof to Theorem 2

Proof: $\{\gamma_t\}$ is a diminishing sequence and $\gamma_t = \frac{\gamma_0}{1+t}$, such that $\sigma \leq K$ and $M_K = KM + K^5 M$. Taking total expectation of (12) in Lemma 1 and summing it from $t = 0$ to $T-1$, we obtain:

$$\mathbb{E}\left[f(w^T)\right] - f(w^0) \leq -\frac{1}{2}\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 + \sum_{t=0}^{T-1}\gamma_t^2 L M_K. \tag{11}$$

Suppose w^* is the optimal solution for $f(w)$, therefore $f(w^*) - f(w^0) \leq \mathbb{E}\left[f(w^T)\right] - f(w^0)$. Letting $\Gamma_T = \sum_{t=0}^{T-1}\gamma_t$, we have:

$$\frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 \leq \frac{2\left(f(w^0) - f(w^*)\right)}{\Gamma_T} + \frac{2\sum_{t=0}^{T-1}\gamma_t^2 L M_K}{\Gamma_T}. \tag{12}$$

We complete the proof.