

## A Convergence Analysis of DaSGD

### A.1 Assumptions

We define some notations.  $\mathcal{S}$  is the training dataset,  $\mathcal{S}_k$  is set  $\{s_k^{(1)}, \dots, s_k^{(M)}\}$  of randomly sampled local batches at  $M$  workers in  $k$  iteration,  $L$  is the Lipschitz constant,  $d$  is the number of local iteration that global weight updates are delayed,  $\tau$  is the number of local steps,  $x$  is the weight of devices. The convergence analysis is conducted under the following assumptions:

- Lipschitzian gradient:  $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$
- Unbiased gradients:  $E_{\mathcal{S}_k|x}[g(x)] = \nabla F(x)$
- Lower bound:  $F(x) \geq F_{inf}$
- Bounded variance:  $E_{\mathcal{S}_k|x}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$
- Independence: All random variables are independent to each other
- Bounded age: The delay is bounded,  $d \leq \tau$

### A.2 Update Rule

The update rule of DaSGD is given by

$$x_{k+1}^{(m)} = \begin{cases} x_k^{(m)} - \eta g(x_k^{(m)}), & \text{otherwise} \\ \xi x_k^{(m)} - \eta \xi g(x_k^{(m)}) + \frac{1-\xi}{M} \sum_{j=1}^M [x_{k-d}^{(j)} - \eta g(x_{k-d}^{(j)})], & (k+1-d) \bmod \tau = 0 \end{cases}$$

where  $x_k^{(m)}$  is the weights at  $m$  worker in  $k$  iteration,  $\eta$  is the learning rate,  $M$  is the number of workers,  $g(x_k^{(m)})$  is the stochastic gradient of worker  $m$ ,  $\xi$  is the local update proportion, delayed update is the case  $(k+1-d) \bmod \tau = 0$ .

**Matrix Representation.** Define matrices  $\mathbf{X}_k, \mathbf{G}_k \in \mathbb{R}^{d \times M}$  that concatenate all local models and gradients in  $k$  iteration:

$$\mathbf{X}_k = [x_k^1, \dots, x_k^M], \mathbf{G}_k = [g(x_k^{(1)}), \dots, g(x_k^{(M)})]$$

Then, the update rule is

$$\mathbf{X}_{k+1} = \begin{cases} \xi(\mathbf{X}_k - \eta \mathbf{G}_k) + (1-\xi)(\mathbf{X}_{k-d} - \eta \mathbf{G}_{k-d}) \mathbf{J}, & (k+1-d) \bmod \tau = 0 \\ \mathbf{X}_k - \eta \mathbf{G}_k, & \text{otherwise} \end{cases} \quad (1)$$

**Update Rule for the Averaged Model.** The update rule of DaSGD is given by

$$x_{k+1}^{(m)} = \begin{cases} x_k^{(m)} - \eta g(x_k^{(m)}), & \text{otherwise} \\ \xi x_k^{(m)} - \eta \xi g(x_k^{(m)}) + \frac{1-\xi}{M} \sum_{j=1}^M [x_{k-d}^{(j)} - \eta g(x_{k-d}^{(j)})], & (k+1-d) \bmod \tau = 0 \end{cases}$$

Here, we set

$$\bar{x}_k = \frac{1}{M} \sum_{i=1}^M x_k^{(i)}, \bar{g}_k = \frac{1}{M} \sum_{i=1}^M g(x_k^{(i)})$$

The average weight on different workers is obtained by

$$\bar{x}_{k+1} = \begin{cases} \bar{x}_k - \eta \bar{g}_k, & \text{otherwise} \\ \xi \bar{x}_k + (1-\xi) \bar{x}_{k-d} - \eta \xi \bar{g}_k - \eta(1-\xi) \bar{g}(x_{k-d}), & (k+1-d) \bmod \tau = 0 \end{cases}$$

When  $z = \tau(k+1)$  for  $z \bmod \tau = 0$ , we have

$$\begin{aligned}
\bar{x}_{\tau(k+1)+d} &= \xi \bar{x}_{\tau(k+1)+d-1} + (1-\xi) \bar{x}_{\tau(k+1)-1} - \xi \eta \bar{g}_{\tau(k+1)+d-1} - (1-\xi) \eta \bar{g}_{\tau(k+1)-1} \\
&= \xi \bar{x}_{\tau k+d} + (1-\xi) \bar{x}_{\tau k+d} - \xi \eta \sum_{i=0}^{\tau-1} \bar{g}_{\tau k+d+i} - (1-\xi) \eta \sum_{i=0}^{\tau-1-d} \bar{g}_{\tau k+d+i} \\
&= \bar{x}_{\tau k+d} - \eta \left[ \xi \left( \sum_{i=0}^{\tau-1} \bar{g}_{\tau k+d+i} - \sum_{i=0}^{\tau-1-d} \bar{g}_{\tau k+d+i} \right) + \sum_{i=0}^{\tau-1-d} \bar{g}_{\tau k+d+i} \right] \\
&= \bar{x}_{\tau k+d} - \eta \left[ \xi \sum_{i=\tau-d}^{\tau-1} \bar{g}_{\tau k+d+i} + \sum_{i=0}^{\tau-1-d} \bar{g}_{\tau k+d+i} \right]
\end{aligned}$$

If we set  $K(k) = \tau k + d$

$$\bar{x}_{K(k+1)} = \bar{x}_{K(k)} - \eta \left[ \xi \sum_{i=\tau-d}^{\tau-1} \bar{g}_{K(k)+i} + \sum_{i=0}^{\tau-1-d} \bar{g}_{K(k)+i} \right]$$

For the ease of writing, we first define some notations. Let  $\mathcal{S}_k$  denote the set  $\{s_k^{(1)}, \dots, s_k^{(m)}\}$  of mini-batches at  $m$  workers in iteration  $k$ . Besides, define averaged stochastic gradient and averaged full batch gradient as follows:

$$\mathcal{G}_{K(k)} = \frac{1}{M} \sum_{m=1}^M \left[ \sum_{i=\tau-d}^{\tau-1} \xi g(x_{\tau k+d+i}^{(m)}) + \sum_{i=0}^{\tau-1-d} g(x_{\tau k+d+i}^{(m)}) \right] \quad (2)$$

$$\mathcal{H}_{K(k)} = \frac{1}{M} \sum_{m=1}^M \left[ \sum_{i=\tau-d}^{\tau-1} \xi \nabla F(x_{\tau k+d+i}^{(m)}) + \sum_{i=0}^{\tau-1-d} \nabla F(x_{\tau k+d+i}^{(m)}) \right] \quad (3)$$

$$\mu_{K(k)} = \frac{1}{M} \sum_{i=1}^M x_{\tau k+d}^{(i)} \quad (4)$$

Then we have

$$\mu_{K(k+1)} = \mu_{K(k)} - \eta \mathcal{G}_{K(k)}$$

### A.3 Convergence Rate

**Theorem (Convergence of DaSGD).** When the learning rate satisfies the following two formulas at the same time

$$\begin{aligned}
2L\eta d\xi^2 - \xi + \frac{6\xi L^2\eta^2 d + 6L^2\eta^2(\tau-d)}{1-\xi^2} + 6\xi L^2\eta^2 d &\leq 0 \\
2L\eta(\tau-d) - \xi + \frac{6\xi L^2\eta^2 d + 6L^2\eta^2(\tau-d)}{1-\xi^2} + 6\xi L^2\eta^2 d + 6L^2\eta^2(\tau-d) &\leq 0
\end{aligned}$$

Then the average-squared gradient norm after  $K$  iterations is bounded as

$$\begin{aligned}
&\mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] \\
&\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{\eta 2L\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} + \eta^2 \frac{6L^2(1+\xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \eta^2 \frac{6d\xi^2 L^2 \tau \sigma^2 (1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} \\
&\quad + \frac{\eta^2}{K} \frac{12d\sigma^2 L^2 \xi^2 (\tau-d)(1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} + \frac{\eta^2}{KM} \frac{12L^2 d \xi^2 (\tau-d)(1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2
\end{aligned}$$

where  $\mu_k = \frac{1}{M} \sum_{i=1}^M x_{\tau k+d}^{(i)}$ ,  $\|\cdot\|_F^2$  is the Frobenius norm.

**Corollary.** Under assumptions, if the learning rate is  $\eta = \frac{M+V}{M} \sqrt{\frac{M}{K}}$  the average-squared gradient norm after  $K$  iterations is bounded by

$$\begin{aligned} & \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] \\ & \leq \frac{2[F(\mu_1) - F_{inf}]}{\sqrt{MK}(\xi d + \tau - d)} + \frac{1}{\sqrt{MK}} \frac{2L\sigma^2 [\xi^2 d + \tau - d]}{(\xi d + \tau - d)} \\ & \quad + \frac{M^2}{K^3} \left(1 + \frac{V}{M}\right)^4 \frac{6L^2(1+\xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{M^2}{K^3} \left(1 + \frac{V}{M}\right)^4 \frac{6d\xi^2 L^2 \tau \sigma^2 (1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} \\ & \quad + \frac{M^2}{K^3} \left(1 + \frac{V}{M}\right)^4 \frac{12d\sigma^2 L^2 \xi^2 (\tau - d)(1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} + \frac{M}{K^3} \left(1 + \frac{V}{M}\right)^4 \frac{12L^2 d \xi^2 (\tau - d)(1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 \end{aligned}$$

If the total iterations  $K$  is sufficiently large, then the average-squared gradient norm will be bounded by

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_k)\|^2 \right] \leq \frac{2[F(\mu_1) - F_{inf}] + 2L\sigma^2 [\xi^2 d + \tau - d]}{\sqrt{MK}(\xi d + \tau - d)}$$

#### A.4 Proof of Convergence Rate

**Lemma 1.** If the learning rate satisfies  $\eta \leq \min \left\{ \frac{1}{2Ld\xi}, \frac{\xi}{2L(\tau-d)} \right\}$  and all local model parameters are initialized at the same point, then the average-squared gradient after  $K$  iterations is bounded as follows

$$\begin{aligned} & \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] \\ & \leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{2L\eta\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} \\ & \quad + \frac{L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \right] \end{aligned}$$

**Proof.**

From the Lipschitzian gradient assumption  $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$ , we have

$$\begin{aligned} F(X_{K(k+1)}) - F(X_{K(k)}) & \leq \langle \nabla F(X_{K(k)}), X_{K(k+1)} - X_{K(k)} \rangle + \frac{L}{2} \|X_{K(k+1)} - X_{K(k)}\|^2 \\ & = -\eta \langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle + \frac{L\eta^2}{2} \|\mathcal{G}_{K(k)}\|^2 \end{aligned} \quad (5)$$

Taking expectation respect to  $\mathcal{S}_{K(k)}$  on both sides of (5), we have

$$\mathbb{E}_{K(k)} [F(X_{K(k+1)})] - F(X_{K(k)}) \leq -\eta \mathbb{E}_{K(k)} [\langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle] + \frac{L\eta^2}{2} \mathbb{E}_{K(k)} [\|\mathcal{G}_{K(k)}\|^2]$$

From the fact

$$\langle a, b \rangle = \frac{1}{2} (\|a\|^2 + \|b\|^2 - \|a - b\|^2)$$

we have

$$\mathbb{E}_{K(k)} [F(X_{K(k+1)})] - F(X_{K(k)}) \leq -\eta \mathbb{E}_{K(k)} [\langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle] + \frac{L\eta^2}{2} \mathbb{E}_{K(k)} [\|\mathcal{G}_{K(k)}\|^2]$$

Combining with Lemmas 4 and 5, we obtain

$$\begin{aligned} & \mathbb{E}_{K(k)} [F(X_{K(k+1)})] - F(X_{K(k)}) \\ & \leq -\eta \mathbb{E}_{K(k)} [\langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle] + \frac{L\eta^2}{2} \mathbb{E}_{K(k)} [\|\mathcal{G}_{K(k)}\|^2] \\ & \leq -\eta \frac{\xi d + \tau - d}{2} \|\nabla F(X_{K(k)})\|^2 + \frac{L\eta^2 \sigma^2}{M} [d\xi^2 + \tau - d] \\ & \quad + \left[ \frac{L\eta^2 d\xi^2}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \left[ \frac{L\eta^2(\tau-d)}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\ & \quad + \frac{\eta}{2M} \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right] \\ & \leq -\eta \frac{\xi d + \tau - d}{2} \|\nabla F(\mu_{K(k)})\|^2 + \frac{L\eta^2 \sigma^2}{M} [d\xi^2 + \tau - d] \\ & \quad + \left[ \frac{L\eta^2 d\xi^2}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \left[ \frac{L\eta^2(\tau-d)}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\ & \quad + \frac{\eta L^2}{2M} \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \right] \tag{6} \end{aligned}$$

where (6) is due to the Lipschitzian gradient assumption  $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$ . After minor rearranging and according to the definition of Frobenius norm, it is easy to show

$$\begin{aligned} & \eta \frac{\xi d + \tau - d}{2} \|\nabla F(\mu_{K(k)})\|^2 \\ & \leq F(\mu_{K(k)}) - \mathbb{E}_{K(k)} [F(\mu_{K(k+1)})] + \frac{L\eta^2 \sigma^2}{M} [d\xi^2 + \tau - d] \\ & \quad + \left[ \frac{L\eta^2 d\xi^2}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \left[ \frac{L\eta^2(\tau-d)}{M} - \frac{\eta\xi}{2M} \right] \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\ & \quad + \frac{\eta L^2}{2M} \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \right] \end{aligned}$$

Taking the total expectation and averaging over all iterates, we have

$$\begin{aligned} & \eta \frac{\xi d + \tau - d}{2} \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] \\ & \leq \frac{F(\mu_1) - F_{inf}}{K} + \frac{L\eta^2 \sigma^2}{M} [d\xi^2 + \tau - d] \\ & \quad + \left[ \frac{L\eta^2 d\xi^2}{KM} - \frac{\eta\xi}{2KM} \right] \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \left[ \frac{L\eta^2(\tau-d)}{KM} - \frac{\eta\xi}{2KM} \right] \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\ & \quad + \frac{\eta L^2}{2KM} \sum_{k=1}^K \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \right] \end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] &\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{2L\eta\sigma^2[\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} \\
&+ \frac{2L\eta d\xi^2 - \xi}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\
&+ \frac{2L\eta(\tau-d) - \xi}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\
&+ \frac{L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \\
&+ \frac{\xi L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2
\end{aligned} \tag{7}$$

If the learning rate satisfies  $\eta \leq \min \left\{ \frac{1}{2Ld\xi}, \frac{\xi}{2L(\tau-d)} \right\}$ , then

$$\begin{aligned}
\mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] &\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{2L\eta\sigma^2[\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} \\
&+ \frac{L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 \\
&+ \frac{\xi L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2
\end{aligned}$$

Recalling the definition  $\mu_{K(k)} = \frac{1}{M} \sum_{i=1}^M x_{\tau k+d}^{(i)} = \mathbf{X}_{K(k)} \mathbf{1}_M / M$  and adding a positive term to the RHS, one can get

$$\sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left\| \mu_{K(k)} - x_{\tau k+d+i}^{(m)} \right\|^2 = \sum_{i=\tau-d}^{\tau-1} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2$$

We have

$$\begin{aligned}
\mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] &\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{2L\eta\sigma^2[\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} \\
&+ \frac{L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2 \\
&+ \frac{\xi L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2
\end{aligned}$$

---

**Lemma 2.**

$$\|\mathcal{H}_{K(k)}\|^2 \leq \frac{2d\xi^2}{M} \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \frac{2(\tau-d)}{M} \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \tag{8}$$

Proof.

$$\begin{aligned}\|\mathcal{H}_{K(k)}\|^2 &= \left\| \xi \frac{1}{M} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \nabla F(x_{\tau k+d+i}^{(m)}) + \frac{1}{M} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &\leq \frac{2d\xi^2}{M^2} \sum_{i=\tau-d}^{\tau-1} \left\| \sum_{m=1}^M \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \frac{2(\tau-d)}{M^2} \sum_{i=0}^{\tau-1-d} \left\| \sum_{m=1}^M \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2\end{aligned}\quad (9)$$

$$\leq \frac{2d\xi^2}{M} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \frac{2(\tau-d)}{M} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2\quad (10)$$

$$= \frac{2d\xi^2}{M} \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \frac{2(\tau-d)}{M} \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2$$

where (9) is due to  $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ , (10) comes from the convexity of vector norm and Jensen's inequality.

**Lemma 3.** Under assumptions  $\mathbb{E}_{\mathcal{S}_k|x}[g(x)] = \nabla F(x)$  and  $\mathbb{E}_{\mathcal{S}_k|x}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ , we have the following variance bound for the averaged stochastic gradient:

$$\mathbb{E}_{K(k)} \left[ \|\mathcal{G}_{K(k)} - \mathcal{H}_{K(k)}\|^2 \right] \leq \frac{2\sigma^2}{M} [d\xi^2 + \tau - d] \quad (11)$$

Proof. According to the definition of (2), (3), and (4), we have

$$\begin{aligned}&\mathbb{E}_{K(k)} \left[ \|\mathcal{G}_{K(k)} - \mathcal{H}_{K(k)}\|^2 \right] \\ &= \frac{1}{M^2} \mathbb{E}_{K(k)} \left[ \left\| \xi \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left[ g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right] + \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \left[ g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right] \right\|^2 \right] \\ &\leq \frac{2}{M^2} \mathbb{E}_{K(k)} \left[ \left\| \xi \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left[ g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right] \right\|^2 + \left\| \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \left[ g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right] \right\|^2 \right]\end{aligned}\quad (12)$$

$$= \frac{2}{M^2} \mathbb{E}_{K(k)} \left[ \xi^2 \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left\| g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \left\| g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right]\quad (13)$$

$$+ \xi^2 \sum_{j \neq i}^{\tau-1} \sum_{l \neq m}^M \left\langle g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}), g(x_{\tau k+d+j}^{(l)}) - \nabla F(x_{\tau k+d+j}^{(l)}) \right\rangle \quad (14)$$

$$+ \sum_{j \neq i}^{\tau-1-d} \sum_{l \neq m}^M \left\langle g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}), g(x_{\tau k+d+j}^{(l)}) - \nabla F(x_{\tau k+d+j}^{(l)}) \right\rangle \quad (15)$$

$$= \frac{2\xi^2}{M^2} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \frac{2}{M^2} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2\quad (16)$$

where (12) is due to  $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ , (16) is due to  $s_k^i$  are independent random variables and the assumption  $\mathbb{E}_{\mathcal{S}_k|x}[g(x)] = \nabla F(x)$ . Now, directly applying assumption  $\mathbb{E}_{\mathcal{S}_k|x}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ ,

$\|\nabla F(x)\|^2 \leq \sigma^2$  to (16). Then, we have

$$\mathbb{E}_{K(k)} \left[ \|\mathcal{G}_{K(k)} - \mathcal{H}_{K(k)}\|^2 \right] \leq \frac{2\xi^2}{M^2} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \sigma^2 + \frac{2}{M^2} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \sigma^2 = \frac{2\sigma^2}{M} [d\xi^2 + \tau - d]$$


---

**Lemma 4.** Under assumption  $\mathbb{E}_{\mathcal{S}_k|x} [g(x)] = \nabla F(x)$ , the expected inner product between stochastic gradient and full batch gradient can be expanded as

$$\begin{aligned} & \mathbb{E}_{K(k)} [\langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle] \\ &= \frac{\xi d + \tau - d}{2} \|\nabla F(X_{K(k)})\|^2 + \frac{1}{2M} \left[ \xi \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \right] \\ & \quad - \frac{1}{2M} \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right] \end{aligned}$$

**Proof.**

$$\begin{aligned} & \mathbb{E}_{K(k)} [\langle \nabla F(X_{K(k)}), \mathcal{G}_{K(k)} \rangle] \\ &= \mathbb{E}_{K(k)} \left[ \left\langle \nabla F(X_{K(k)}), \xi \frac{1}{M} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M g(x_{\tau k+d+i}^{(m)}) + \frac{1}{M} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M g(x_{\tau k+d+i}^{(m)}) \right\rangle \right] \\ &= \xi \frac{1}{M} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \langle \nabla F(X_{K(k)}), \nabla F(x_{\tau k+d+i}^{(m)}) \rangle + \frac{1}{M} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \langle \nabla F(X_{K(k)}), \nabla F(x_{\tau k+d+i}^{(m)}) \rangle \\ &= \frac{\xi}{2M} \sum_{i=\tau-d}^{\tau-1} \sum_{m=1}^M \left[ \|\nabla F(X_{K(k)})\|^2 + \|\nabla F(x_{\tau k+d+i}^{(m)})\|^2 - \|\nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)})\|^2 \right] \end{aligned} \tag{17}$$

$$\begin{aligned} & + \frac{1}{2M} \sum_{i=0}^{\tau-1-d} \sum_{m=1}^M \left[ \|\nabla F(X_{K(k)})\|^2 + \|\nabla F(x_{\tau k+d+i}^{(m)})\|^2 - \|\nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)})\|^2 \right] \\ &= \frac{\xi d + \tau - d}{2} \|\nabla F(X_{K(k)})\|^2 + \frac{1}{2M} \left[ \xi \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \right] \\ & \quad - \frac{1}{2M} \sum_{m=1}^M \left[ \sum_{i=0}^{\tau-1-d} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \xi \sum_{i=\tau-d}^{\tau-1} \left\| \nabla F(X_{K(k)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right] \end{aligned} \tag{18}$$

where (17) and (18) come from  $\langle a, b \rangle = \frac{1}{2} (\|a\|^2 + \|b\|^2 - \|a - b\|^2)$ .

---

**Lemma 5.** Under assumptions  $E_{\xi|x} [g(x)] = \nabla F(x)$  and  $E_{\xi|x} \|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ , the squared norm of stochastic gradient can be bounded as

$$\mathbb{E}_{K(k)} \left[ \|\mathcal{G}_{K(k)}\|^2 \right] \leq \frac{2\sigma^2}{M} [d\xi^2 + \tau - d] + \frac{2d\xi^2}{M} \sum_{i=\tau-d}^{\tau-1} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + \frac{2(\tau-d)}{M} \sum_{i=0}^{\tau-1-d} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2$$

Proof.

$$\begin{aligned}
& \mathbb{E}_{K(k)} \left[ \left\| \mathcal{G}_{K(k)} \right\|^2 \right] \\
&= \mathbb{E}_{K(k)} \left[ \left\| \mathcal{G}_{K(k)} - \mathbb{E}_{K(k)}[\mathcal{G}_{K(k)}] \right\|^2 \right] + \left\| \mathbb{E}_{K(k)}[\mathcal{G}_{K(k)}] \right\|^2 \\
&= \mathbb{E}_{K(k)} \left[ \left\| \mathcal{G}_{K(k)} - \mathcal{H}_{K(k)} \right\|^2 \right] + \left\| \mathcal{H}_{K(k)} \right\|^2 \\
&\leq \frac{2\sigma^2}{M} [d\xi^2 + \tau - d] + \frac{2d\xi^2}{M} \sum_{i=\tau-d}^{\tau-1} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|_F^2 + \frac{2(\tau-d)}{M} \sum_{i=0}^{\tau-1-d} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|_F^2
\end{aligned} \tag{19}$$

where (19) follows (8) and (11).

**Theorem 1 (Convergence of SGD).** Under assumptions, when the learning rate satisfies the following two formulas at the same time

$$\begin{aligned}
2L\eta d\xi^2 - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau - d)}{1 - \xi^2} + 6\xi L^2 \eta^2 d &\leq 0 \\
2L\eta(\tau - d) - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau - d)}{1 - \xi^2} + 6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau - d) &\leq 0
\end{aligned}$$

Then the average-squared gradient norm after  $K$  iterations is bounded as

$$\begin{aligned}
& \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \nabla F(\mu_{K(k)}) \right\|^2 \right] \\
&\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{\eta 2L\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} + \eta^2 \frac{6L^2(1 + \xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \eta^2 \frac{6d\xi^2 L^2 \tau \sigma^2 (1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \\
&\quad + \frac{\eta^2}{K} \frac{12d\sigma^2 L^2 \xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} + \frac{\eta^2}{KM} \frac{12L^2 d\xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \sum_{i=1}^{d-1} \left\| \nabla F(\mathbf{X}_i) \right\|_F^2
\end{aligned}$$

where  $\mu_k = \frac{1}{M} \sum_{i=1}^M x_{\tau k+d}^{(i)}$ ,  $\left\| \cdot \right\|_F$  is the Frobenius norm.

Proof.

Recall the intermediate result (7) in the proof of Lemma 1:

$$\begin{aligned}
\mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \nabla F(\mu_{K(k)}) \right\|^2 \right] &\leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{2L\eta\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} \\
&\quad + \frac{2L\eta d\xi^2 - \xi}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|_F^2 \\
&\quad + \frac{2L\eta(\tau - d) - \xi}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|_F^2 \\
&\quad + \frac{L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \left\| \mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i} \right\|_F^2 \\
&\quad + \frac{\xi L^2}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i} \right\|_F^2
\end{aligned} \tag{20}$$

Our goal is to provide an upper bound for the network error term  $\sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i} \right\|_F^2$ . First of all, let us derive a specific expression for



$\mathbf{X}_{\tau k+d}\mathbf{J} - \mathbf{X}_{\tau k+d+i}$ . According to the update rule (1), one can observe that

$$\begin{aligned}
& \mathbf{X}_{\tau k+d}\mathbf{J} - \mathbf{X}_{\tau k+d+i} \\
&= \mathbf{X}_{\tau k+d}(\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi (\mathbf{X}_{\tau k+d-1} - \eta \mathbf{G}_{\tau k+d-1}) (\mathbf{J} - \mathbf{I}) + (1 - \xi) (\mathbf{X}_{\tau k} - \eta \mathbf{G}_{\tau k}) \mathbf{J} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi \mathbf{X}_{\tau(k-1)+d} (\mathbf{J} - \mathbf{I}) - \xi \eta \sum_{i=0}^{\tau-1} \mathbf{G}_{\tau(k-1)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi^2 \mathbf{X}_{\tau(k-2)+d} (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^2 \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi^k \mathbf{X}_d (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi^k (\mathbf{X}_{d-1} - \eta \mathbf{G}_{d-1}) (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= \xi^k \mathbf{X}_1 (\mathbf{J} - \mathbf{I}) - \eta \xi^k \sum_{i=1}^{d-1} \mathbf{G}_i (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \\
&= -\eta \xi^k \sum_{i=1}^{d-1} \mathbf{G}_i (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{j=0}^i \mathbf{G}_{\tau k+d+j} \tag{21}
\end{aligned}$$

where (21) follows the fact that all workers start from the same point at the beginning of each local update period.

Accordingly, we have

$$\begin{aligned}
& \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2 \\
&= \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| -\eta \xi^k \sum_{i=1}^{d-1} \mathbf{G}_i (\mathbf{J} - \mathbf{I}) - \eta \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) + \eta \sum_{i=0}^l \mathbf{G}_{\tau k+d+i} \right\|_F^2 \\
&\leq 3\eta^2 \mathbb{E}_{K(k)} \left[ \xi^{2k} d \left\| \sum_{i=1}^{d-1} \mathbf{G}_i (\mathbf{J} - \mathbf{I}) \right\|_F^2 + d \left\| \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} (\mathbf{J} - \mathbf{I}) \right\|_F^2 + \sum_{l=\tau-d}^{\tau-1} \left\| \sum_{i=0}^l \mathbf{G}_{\tau k+d+i} \right\|_F^2 \right] \\
&\leq 3\eta^2 \mathbb{E}_{K(k)} \left[ \xi^{2k} d \left\| \sum_{i=1}^{d-1} \mathbf{G}_i \right\|_F^2 + d \left\| \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j \mathbf{G}_{\tau(k-j)+d+i} \right\|_F^2 + \sum_{l=\tau-d}^{\tau-1} \left\| \sum_{i=0}^l \mathbf{G}_{\tau k+d+i} \right\|_F^2 \right] \quad (22) \\
&= 3\eta^2 \sum_{m=1}^M \left[ \mathbb{E}_{K(k)} \xi^{2k} d \left\| \sum_{i=1}^{d-1} g(x_i^{(m)}) \right\|^2 + d \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j g(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2 + \mathbb{E}_{K(k)} \sum_{l=\tau-d}^{\tau-1} \left\| \sum_{i=0}^l g(x_{\tau k+d+i}^{(m)}) \right\|^2 \right] \\
&= 3\eta^2 d \left[ \underbrace{\sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} g(x_i^{(m)}) \right\|^2}_{T_1} + \underbrace{\sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \sum_{i=0}^{\tau-1} \xi^j g(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2}_{T_2} + \underbrace{\frac{1}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \sum_{i=0}^l g(x_{\tau k+d+i}^{(m)}) \right\|^2}_{T_3} \right] \quad (23)
\end{aligned}$$

where the (22) is due to the operator norm of  $\mathbf{J} - \mathbf{I}$  is less than 1.

For  $T_2$ , we have

$$\begin{aligned}
& \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \xi^j \sum_{i=0}^{\tau-1} g(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2 \\
&= \sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \xi^j \sum_{i=0}^{\tau-1} [g(x_{\tau(k-j)+d+i}^{(m)}) - \nabla F(x_{\tau(k-j)+d+i}^{(m)})] + \sum_{j=1}^k \xi^j \sum_{i=0}^{\tau-1} \nabla F(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2 \\
&\leq 2 \underbrace{\sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \xi^j \sum_{i=0}^{\tau-1} [g(x_{\tau(k-j)+d+i}^{(m)}) - \nabla F(x_{\tau(k-j)+d+i}^{(m)})] \right\|^2}_{T_4} + 2 \underbrace{\sum_{m=1}^M \mathbb{E}_{K(k)} \left\| \sum_{j=1}^k \xi^j \sum_{i=0}^{\tau-1} \nabla F(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2}_{T_5}
\end{aligned}$$

For the first term  $T_4$ , since the stochastic gradients are unbiased, all cross terms are zero. Thus, combining with Assumption of bounded variance, we have

$$\begin{aligned}
T_4 &= 2 \sum_{m=1}^M \sum_{j=1}^k \xi^{2j} \sum_{i=0}^{\tau-1} \mathbb{E}_{K(k)} \left\| g(x_{\tau(k-j)+d+i}^{(m)}) - \nabla F(x_{\tau(k-j)+d+i}^{(m)}) \right\|^2 \\
&\leq 2 \sum_{m=1}^M \sum_{j=1}^k \xi^{2j} \sum_{i=0}^{\tau-1} \sigma^2 \leq \frac{2M\tau\sigma^2\xi^2}{1-\xi^2} \quad (24)
\end{aligned}$$

where (24) according to the summation formula of power

$$\sum_{j=1}^k \xi^{2j} \leq \sum_{j=1}^{\infty} \xi^{2j} \leq \frac{\xi^2}{1-\xi^2}$$

For the second term  $T_5$ , we get

$$T_5 = 2 \sum_{j=1}^k \xi^{2j} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau(k-j)+d+i}^{(m)}) \right\|_F^2 = 2 \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2$$

Substituting the bounds of  $T_4$  and  $T_5$  into  $T_2$ , we have

$$T_2 \leq \frac{2M\tau\sigma^2\xi^2}{1-\xi^2} + 2 \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2$$

For  $T_1$ , we have

$$\begin{aligned} & \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} g(x_i^{(m)}) \right\|^2 \\ &= \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} [g(x_i^{(m)}) - \nabla F(x_i^{(m)})] + \sum_{i=1}^{d-1} \nabla F(x_i^{(m)}) \right\|^2 \\ &\leq \underbrace{2 \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} [g(x_i^{(m)}) - \nabla F(x_i^{(m)})] \right\|^2}_{T_6} + \underbrace{2 \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} \nabla F(x_i^{(m)}) \right\|^2}_{T_7} \end{aligned}$$

For the first term  $T_6$ , since the stochastic gradients are unbiased, all cross terms are zero. Thus, combining with Assumption of bounded variance, we have

$$T_6 = 2 \sum_{m=1}^M \sum_{i=1}^{d-1} \mathbb{E}_{K(k)} \xi^{2k} \left\| g(x_i^{(m)}) - \nabla F(x_i^{(m)}) \right\|^2 \leq 2\xi^{2k} \sum_{m=1}^M \sum_{i=1}^{d-1} \sigma^2 = 2\xi^{2k} M d \sigma^2$$

For the second term  $T_7$ , directly applying Jensen's inequality, we get

$$T_7 = 2 \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \left\| \sum_{i=1}^{d-1} \nabla F(x_i^{(m)}) \right\|^2 \leq 2d \sum_{m=1}^M \mathbb{E}_{K(k)} \xi^{2k} \sum_{i=1}^{d-1} \left\| \nabla F(x_i^{(m)}) \right\|^2 = 2d\xi^{2k} \sum_{i=1}^{d-1} \left\| \nabla F(\mathbf{X}_i) \right\|_F^2$$

Substituting the bounds of  $T_6$  and  $T_7$  into  $T_1$ , we have

$$T_1 \leq 2\xi^{2k} M d \sigma^2 + 2d\xi^{2k} \sum_{i=1}^{d-1} \left\| \nabla F(\mathbf{X}_i) \right\|_F^2$$

For  $T_3$ , we have

$$\begin{aligned} T_3 &= \frac{1}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \sum_{i=0}^l g(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &= \frac{1}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \sum_{i=0}^l (g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)})) + \sum_{i=0}^l \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &\leq \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \sum_{i=0}^l (g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)})) \right\|^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \left\| \sum_{i=0}^l \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &\leq \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| g(x_{\tau k+d+i}^{(m)}) - \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &\leq \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \\ &= \frac{2M}{d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \end{aligned}$$

Substituting the bounds of  $T_1$ ,  $T_2$  and  $T_3$  into (23), we have

$$\begin{aligned}
& \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2 \\
& \leq 3\eta^2 d \left[ 2\xi^{2k} M d \sigma^2 + \frac{2M\tau\sigma^2\xi^2}{1-\xi^2} + 2d\xi^{2k} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 + 2 \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2 \right. \\
& \quad \left. + \frac{2M}{d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right]
\end{aligned}$$

And in the same way, we have

$$\begin{aligned}
& \sum_{l=0}^{\tau-1-d} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+l}\|_F^2 \\
& \leq 3\eta^2 (\tau-d) \left[ 2\xi^{2k} M d \sigma^2 + \frac{2M\tau\sigma^2\xi^2}{1-\xi^2} + 2d\xi^{2k} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 + 2 \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2 \right. \\
& \quad \left. + \frac{2M}{\tau-d} \sum_{l=0}^{\tau-1-d} \sum_{i=0}^l \sigma^2 + \frac{2}{\tau-d} \sum_{m=1}^M \sum_{l=0}^{\tau-1-d} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right]
\end{aligned}$$

Then, summing over all periods from  $k = 0$  to  $k = K$ , where  $K$  is the total global iterations:

$$\begin{aligned}
& \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2 \\
& \leq 3\eta^2 d \sum_{k=1}^K \left[ 2\xi^{2k} M d \sigma^2 + \frac{2M\tau\sigma^2\xi^2}{1-\xi^2} + 2d\xi^{2k} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 + 2 \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2 \right. \\
& \quad \left. + \frac{2M}{d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{2}{d} \sum_{m=1}^M \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(x_{\tau k+d+i}^{(m)}) \right\|^2 \right] \\
& \leq 6\eta^2 d \frac{\xi^2}{1-\xi^2} \left[ 2M d \sigma^2 + 2d \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 \right] + \frac{6\eta^2 d \tau \sigma^2 \xi^2 M K}{1-\xi^2} + 6\eta^2 M K \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 \\
& \quad + 6\eta^2 d \sum_{k=1}^K \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2 + 6\eta^2 \sum_{k=1}^K \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E}_{K(k)} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|^2
\end{aligned} \tag{25}$$

Expanding the summation, we have

$$\begin{aligned}
& \sum_{k=1}^K \sum_{r=0}^{k-1} \xi^{2(k-r)} \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}^{(m)}) \right\|_F^2 \\
& \leq \sum_{r=1}^K \left[ \left( \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}) \right\|_F^2 \right) \left( \sum_{k=r}^K \xi^{2(k-r)} \right) \right] \\
& \leq \sum_{r=1}^K \left[ \left( \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau r+d+i}) \right\|_F^2 \right) \left( \sum_{k=r}^{+\infty} \xi^{2(k-r)} \right) \right] \\
& \leq \frac{1}{1-\xi^2} \sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E} \left\| \nabla F(\mathbf{X}_{\tau k+d+i}) \right\|_F^2
\end{aligned} \tag{26}$$

And in the same way, we have

$$\sum_{k=1}^K \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \leq d \sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \quad (27)$$

Plugging (26) and (27) into (25),

$$\begin{aligned} & \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+i}\|_F^2 \\ & \leq 6\eta^2 d \frac{\xi^2}{1-\xi^2} \left[ 2Md\sigma^2 + 2d \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 \right] + \frac{6\eta^2 d \tau \sigma^2 \xi^2 MK}{1-\xi^2} + 6\eta^2 MK \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 \\ & \quad + \frac{6\eta^2 d}{1-\xi^2} \sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + 6\eta^2 d \sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \end{aligned}$$

And in the same way, we have

$$\begin{aligned} & \sum_{k=1}^K \sum_{l=0}^{\tau-1-d} \mathbb{E}_{K(k)} \|\mathbf{X}_{\tau k+d} \mathbf{J} - \mathbf{X}_{\tau k+d+l}\|_F^2 \\ & \leq 6\eta^2 (\tau-d) \frac{\xi^2}{1-\xi^2} \left[ 2Md\sigma^2 + 2d \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 \right] + \frac{6\eta^2 (\tau-d) \tau \sigma^2 \xi^2 MK}{1-\xi^2} + 6\eta^2 MK \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 \\ & \quad + \frac{6\eta^2 (\tau-d)}{1-\xi^2} \sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 + 6\eta^2 (\tau-d) \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \end{aligned}$$

Recall the intermediate result (7) in the proof of Lemma 1:

$$\begin{aligned} & \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(\mu_{K(k)})\|^2 \right] \\ & \leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{\eta 2L\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} + \eta^2 \frac{6L^2(1+\xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \eta^2 \frac{6d\xi^2 L^2 \tau \sigma^2 (1+\xi)}{(\xi d + \tau - d)(1-\xi^2)} \\ & \quad + \frac{\eta^2 12d\sigma^2 L^2 \xi^2 (\tau-d)(1+\xi)}{K(\xi d + \tau - d)(1-\xi^2)} + \frac{\eta^2 12L^2 d \xi^2 (\tau-d)(1+\xi)}{KM(\xi d + \tau - d)(1-\xi^2)} \sum_{i=1}^{d-1} \|\nabla F(\mathbf{X}_i)\|_F^2 \\ & \quad + \frac{2L\eta d \xi^2 - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d)}{1-\xi^2} + 6\xi L^2 \eta^2 d}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=\tau-d}^{\tau-1} \mathbb{E}_{K(k)} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \\ & \quad + \frac{2L\eta(\tau-d) - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d)}{1-\xi^2} + 6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d)}{KM(\xi d + \tau - d)} \sum_{k=1}^K \sum_{i=0}^{\tau-1-d} \mathbb{E}_{K(k)} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 \end{aligned}$$

When the learning rate satisfies the following two formulas at the same time

$$\begin{aligned} & 2L\eta d \xi^2 - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d)}{1-\xi^2} + 6\xi L^2 \eta^2 d \leq 0 \\ & 2L\eta(\tau-d) - \xi + \frac{6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d)}{1-\xi^2} + 6\xi L^2 \eta^2 d + 6L^2 \eta^2 (\tau-d) \leq 0 \end{aligned}$$

And

$$\sum_{k=1}^K \sum_{i=0}^{\tau-1} \mathbb{E}_{K(k)} \|\nabla F(\mathbf{X}_{\tau k+d+i})\|_F^2 = \mathbb{E}_k \sum_{k=1}^K \|\nabla F(\mu_k)\|^2$$

Thus, we have

$$\begin{aligned}
& \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \nabla F(\mu_{K(k)}) \right\|^2 \right] \\
& \leq \frac{2[F(\mu_1) - F_{inf}]}{\eta K(\xi d + \tau - d)} + \frac{\eta 2L\sigma^2 [\xi^2 d + \tau - d]}{M(\xi d + \tau - d)} + \eta^2 \frac{6L^2(1 + \xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \eta^2 \frac{6d\xi^2 L^2 \tau \sigma^2 (1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \\
& \quad + \frac{\eta^2}{K} \frac{12d\sigma^2 L^2 \xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} + \frac{\eta^2}{KM} \frac{12L^2 d \xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \sum_{i=1}^{d-1} \left\| \nabla F(\mathbf{X}_i) \right\|_F^2
\end{aligned}$$


---

**Corollary 1.** Under assumptions, if the learning rate is  $\eta = \frac{M+V}{M} \sqrt{\frac{M}{K}}$  the average-squared gradient norm after  $K$  iterations is bounded by

$$\begin{aligned}
& \mathbb{E}_{K(k)} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \nabla F(\mu_{K(k)}) \right\|^2 \right] \\
& \leq \frac{2[F(\mu_1) - F_{inf}]}{\sqrt{MK}(\xi d + \tau - d)} + \frac{1}{\sqrt{MK}} \frac{2L\sigma^2 [\xi^2 d + \tau - d]}{(\xi d + \tau - d)} \\
& \quad + \frac{M^2}{K^3} \left( 1 + \frac{V}{M} \right)^4 \frac{6L^2(1 + \xi)}{\xi d + \tau - d} \sum_{l=\tau-d}^{\tau-1} \sum_{i=0}^l \sigma^2 + \frac{M^2}{K^3} \left( 1 + \frac{V}{M} \right)^4 \frac{6d\xi^2 L^2 \tau \sigma^2 (1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \\
& \quad + \frac{M^2}{K^3} \left( 1 + \frac{V}{M} \right)^4 \frac{12d\sigma^2 L^2 \xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} + \frac{M}{K^3} \left( 1 + \frac{V}{M} \right)^4 \frac{12L^2 d \xi^2 (\tau - d)(1 + \xi)}{(\xi d + \tau - d)(1 - \xi^2)} \sum_{i=1}^{d-1} \left\| \nabla F(\mathbf{X}_i) \right\|_F^2
\end{aligned}$$

If the total iterations  $K$  is sufficiently large, then the average-squared gradient norm will be bounded by

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^K \left\| \nabla F(\mu_k) \right\|^2 \right] \leq \frac{2[F(\mu_1) - F_{inf}] + 2L\sigma^2 [\xi^2 d + \tau - d]}{\sqrt{MK}(\xi d + \tau - d)}$$