Online Supplement

Tackling Arbitrarily Heterogeneous Data in Asynchronous SGD without Worker Scheduling

Xiaolu Wang   Yuchang Sun   Hoi-To Wai   Jun Zhang

1 Notation for Conditional Expectations

For notational simplicity, we define

$$\mathbb{E}_P[h(P,Q)] \coloneqq \mathbb{E}[h(P,Q) \mid Q]$$

for random variables P, Q and a measurable function h, which represents the *conditional expectation* with respect to P while holding Q fixed.

Throughout the Appendix, we frequently encounter terms of the form $\mathbb{E}_{\boldsymbol{\xi}_i^{t-\rho_i(t)}}\left[\nabla f_i\left(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}\right)\right]$ with $t-\rho_i(t) > t-\tau_i(t)$. By definition, this denotes the conditional expectation over the sample $\boldsymbol{\xi}_i^{t-\rho_i(t)}$ given the current iterate $\boldsymbol{w}^{t-\tau_i(t)}$ (equivalently, given the sigma algebra $\mathcal{F}_{t-\tau_i(t)}$ generated by the history up to time $t-\tau_i(t)$). Since $\boldsymbol{w}^{t-\tau_i(t)}$ is $\mathcal{F}_{t-\tau_i(t)}$-measurable, we have

$$\mathbb{E}_{\boldsymbol{\xi}_i^{t-\rho_i(t)}}\left[\nabla f_i\left(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}\right)\right] = \mathbb{E}\left[\nabla f_i\left(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}\right) \,\middle|\, \mathcal{F}_{t-\tau_i(t)}\right] = \nabla F_i\left(\boldsymbol{w}^{t-\tau_i(t)}\right),$$

where the last equality follows from the conditional unbiasedness assumption (Assumption 3) and the fact that $\boldsymbol{\xi}_i^{t-\rho_i(t)}$ is sampled independently of $\mathcal{F}_{t-\tau_i(t)}$.

2 Technical Lemmas

Lemma S.1. *Suppose that Assumptions 3 and 4 hold. Then, it holds for all $t \geq 1$ that*

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right)\right\|_2^2 \leq \frac{\sigma^2}{n}.$$

Proof. Expanding the squared norm gives

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right)\right\|_2^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\|_2^2 \\
&\quad + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\mathbb{E}\left\langle \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}), \nabla f_j(\boldsymbol{w}^{t-\tau_j(t)}; \boldsymbol{\xi}_j^{t-\rho_j(t)}) - \nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\rangle}_{Y_{ij}}
\end{aligned} \tag{S.1}$$

To simplify Y_{ij} for $i, j \in [n]$ such that $i \neq j$, we assume without loss of generality that $\rho_i(t) \geq \rho_j(t)$. Then $t - \tau_i(t) \leq t - \rho_j(t)$ and thus $\boldsymbol{\xi}_j^{t-\rho_j(t)}$ is independent of $\boldsymbol{w}^{t-\tau_i(t)}$. Hence, using the law of total expectation and Assumption 3, we have

$$\begin{aligned}
Y_{ij} &= \mathbb{E}\left\langle \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}), \nabla f_j(\boldsymbol{w}^{t-\tau_j(t)}; \boldsymbol{\xi}_j^{t-\rho_j(t)}) - \nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\rangle \\
&= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\xi}_j^{t-\rho_j(t)}}\left\langle \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}), \nabla f_j(\boldsymbol{w}^{t-\tau_j(t)}; \boldsymbol{\xi}_j^{t-\rho_j(t)}) - \nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}), \mathbb{E}_{\boldsymbol{\xi}_j^{t-\rho_j(t)}}\left[\nabla f_j(\boldsymbol{w}^{t-\tau_j(t)}; \boldsymbol{\xi}_j^{t-\rho_j(t)}) - \nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right]\right\rangle\right] \\
&= 0.
\end{aligned}$$

Substituting this into (S.1) gives

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right)\right\|_2^2 \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\|_2^2 \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\|\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\|_2^2 \,\middle|\, \boldsymbol{w}^{t-\tau_i(t)}\right]\right] \\
\leq& \frac{\sigma^2}{n}. \qquad \text{(S.2)}
\end{aligned}$$

where the inequality follows from Assumption 4. □

Lemma S.2. *Suppose that Assumptions 3 and 4 hold. Then, it holds for all $i \in [n]$ and $t \geq 1$ that*

$$\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t-\tau_i(t)}\|_2^2 \leq 2\tau_{\max}^2\eta^2\frac{\sigma^2}{n} + 2\tau_{\max}\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2.$$

In addition, we have

$$\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{[t-\tau_{\max}]_+}\|_2^2 \leq 2\tau_{\max}^2\eta^2\frac{\sigma^2}{n} + 2\tau_{\max}\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2. \qquad \text{(S.3)}$$

Proof. For all $i \in [n]$ and $t \geq 1$, it follows from the telescoping sum $\sum_{s=1+t-\tau_i(t)}^{t}\left(\boldsymbol{w}^s - \boldsymbol{w}^{s-1}\right) = \boldsymbol{w}^t - \boldsymbol{w}^{t-\tau_i(t)}$ and the iterative formula (3) that

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t-\tau_i(t)}\|_2^2 \\
=& \mathbb{E}\left\|\sum_{s=1+t-\tau_i(t)}^{t}\left(\boldsymbol{w}^s - \boldsymbol{w}^{s-1}\right)\right\|_2^2
\end{aligned}$$

$$= \mathbb{E}\left\| \sum_{s=1+t-\tau_i(t)}^{t} \eta \boldsymbol{g}^s \right\|_2^2$$

$$= \mathbb{E}\left\| \sum_{s=1+t-\tau_i(t)}^{t} \frac{\eta}{n} \sum_{j=1}^{n} \nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) \right\|_2^2$$

$$= \frac{\eta^2}{n^2}\mathbb{E}\left\| \sum_{s=1+t-\tau_i(t)}^{t} \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) + \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right) \right\|_2^2. \tag{S.4}$$

Further applying the fact that $\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 \le 2\|\boldsymbol{x}\|_2^2 + 2\|\boldsymbol{y}\|_2^2$ for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to (S.4), we have

$$\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t-\tau_i(t)}\|_2^2 \le \frac{2\eta^2}{n^2} \underbrace{\mathbb{E}\left\| \sum_{s=1+t-\tau_i(t)}^{t} \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right) \right\|_2^2}_{\Phi_1}$$

$$+ \frac{2\eta^2}{n^2} \underbrace{\mathbb{E}\left\| \sum_{s=1+t-\tau_i(t)}^{t} \sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right\|_2^2}_{\Phi_2}. \tag{S.5}$$

Subsequently, we upper bound Φ_1 and Φ_2, respectively. Expanding Φ_1, we have

$$\Phi_1 = \sum_{s=1+t-\tau_i(t)}^{t} \mathbb{E}\left\| \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_i^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right) \right\|_2^2$$

$$+ \sum_{\substack{s,s':s\neq s',\\ 1+t-\tau_i(t)\le s,s'\le t}} \mathbb{E}\left\langle \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right), \right.$$

$$\left. \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')}) \right) \right\rangle. \tag{S.6}$$

For the inner product terms in (S.6), denoted by $Z^{ss'}$ for $s, s' \in [1+t-\tau_i(t), t]$ and $s \neq s'$, note that

$$Z^{ss'} = \mathbb{E}\left\langle \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right), \sum_{j=1}^{n} \left(\nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')}) \right) \right\rangle$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle \nabla f_i(\boldsymbol{w}^{s-\tau_i(s)}; \boldsymbol{\xi}_i^{s-\rho_i(s)}) - \nabla F_i(\boldsymbol{w}^{s-\tau_i(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')}) \right\rangle \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\left\langle \nabla f_i(\boldsymbol{w}^{s-\tau_i(s)}; \boldsymbol{\xi}_i^{s-\rho_i(s)}) - \nabla F_i(\boldsymbol{w}^{s-\tau_i(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')}) \right\rangle$$

$$= \sum_{j=1}^{n} \mathbb{E}\left\langle \nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{s-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')})\right\rangle$$

$$+ \sum_{i,j:i\neq j} \mathbb{E}\left\langle \nabla f_i(\boldsymbol{w}^{s-\tau_i(s)}; \boldsymbol{\xi}_i^{s-\rho_i(s)}) - \nabla F_i(\boldsymbol{w}^{s-\tau_i(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_j(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_j(s')})\right\rangle.$$

To simplify the inner product terms in the last line of the above equality, denoted by $Z_{ij}^{ss'}$ for all $i, j \in [n]$ and $i \neq j$, we assume without loss of generality that $s - \rho_i(s) > s' - \rho_j(s')$, which implies that $\xi_i^{s-\rho_i(s)}$ is sampled after $\xi_j^{s'-\rho_j(s')}$. Additionally, since $\rho_j(s') \leq \tau_j(s')$ by (5), we have $s - \rho_i(s) > s' - \tau_j(s')$. Thus, $\boldsymbol{\xi}_i^{s-\rho_i(s)}$ is independent of $\boldsymbol{w}^{s'-\rho_j(s')}$. Further using the law of total expectation and Assumption 3, we have

$$\begin{aligned}
Z_{ij}^{ss'} &= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\xi}_i^{s-\rho_i(s)}}\left\langle \nabla f_i(\boldsymbol{w}^{s-\tau_i(s)}; \boldsymbol{\xi}_j^{s-\rho_i(s)}) - \nabla F_i(\boldsymbol{w}^{s-\tau_i(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_i(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_i(s')})\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \mathbb{E}_{\boldsymbol{\xi}_i^{s-\rho_i(s)}}\left[\nabla f_i(\boldsymbol{w}^{s-\tau_i(s)}; \boldsymbol{\xi}_i^{s-\rho_i(s)}) - \nabla F_i(\boldsymbol{w}^{s-\tau_i(s)})\right], \nabla f_j(\boldsymbol{w}^{s'-\tau_i(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_i(s')})\right\rangle\right] \\
&= 0
\end{aligned}$$

Substituting this into $Z^{ss'}$ and using Assumption 4, we have

$$\begin{aligned}
Z^{ss'} &= \sum_{j=1}^{n} \mathbb{E}\left\langle \nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{t-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}), \nabla f_j(\boldsymbol{w}^{s'-\tau_i(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_i(s')})\right\rangle \\
&\leq \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\|\nabla f_j(\boldsymbol{w}^{s-\tau_j(s)}; \boldsymbol{\xi}_j^{t-\rho_j(s)}) - \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\|_2^2 \\
&\quad + \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\|\nabla f_j(\boldsymbol{w}^{s'-\tau_i(s')}; \boldsymbol{\xi}_j^{s'-\rho_j(s')}) - \nabla F_j(\boldsymbol{w}^{s'-\tau_i(s')})\|_2^2 \\
&\leq n\sigma^2.
\end{aligned}$$

Plugging this back into (S.6) and using Lemma S.1 yield

$$\begin{aligned}
\Phi_1 &\leq \sum_{s=1+t-\tau_i(t)}^{t} n\sigma^2 + \sum_{\substack{s,s':s\neq s',\\ 1+t-\tau_i(t)\leq s,s'\leq t}} n\sigma^2 \\
&= \tau_i(t)n\sigma^2 + (\tau_i(t)^2 - \tau_i(t))n\sigma^2 \\
&= \tau_i(t)^2 n\sigma^2 \\
&\leq n\tau_{\max}^2\sigma^2.
\end{aligned} \tag{S.7}$$

To upper bound Φ_2, we use the fact that $\|\sum_{i=1}^{m} \boldsymbol{x}_i\|_2^2 \leq m\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2$ for vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, then

$$\Phi_2 = \mathbb{E}\left\|\sum_{s=1+t-\tau_i(t)}^{t}\sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2$$

$$\leq \tau_i(t) \sum_{s=1+t-\tau_i(t)}^{t} \mathbb{E}\left\| \sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right\|_2^2$$

$$\leq \tau_{\max} \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\left\| \sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right\|_2^2 . \tag{S.8}$$

Substituting (S.7) and (S.8) back into (S.5) gives the desired result. Additionally, (S.3) can be proved in a similar manner. □

Lemma S.3. *Suppose that Assumptions 2–4 hold. Then, it holds for all $i \in [n]$ and $t \geq 1$ that*

$$\mathbb{E}\|\boldsymbol{g}^t\|_2^2 \leq \left(2 + 8L^2\tau_{\max}^2\eta^2\right)\frac{\sigma^2}{n} + 4\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 + 8L^2\tau_{\max}\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\left\| \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right\|_2^2 .$$

Proof. Following the fact that $\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 \leq 2\|\boldsymbol{x}\|_2^2 + 2\|\boldsymbol{y}\|_2^2$ for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}^t\|_2^2 &= \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) \right\|_2^2 \\
&= \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right) + \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) \right\|_2^2 \\
&\leq 2\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right) \right\|_2^2 + 2\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) \right\|_2^2 \\
&\leq \frac{2\sigma^2}{n} + 2\underbrace{\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) \right\|_2^2}_{\Psi} .
\end{aligned} \tag{S.9}$$

where the last inequality holds due to Lemma S.1. It suffices to upper bound Ψ. We observe that

$$\begin{aligned}
\Psi &\leq 2\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-1})\right) \right\|_2^2 + 2\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-1}) \right\|_2^2 \\
&\leq \frac{2L^2}{n}\sum_{i=1}^{n} \mathbb{E}\|\boldsymbol{w}^{t-\tau_i(t)} - \boldsymbol{w}^{t-1}\|_2^2 + 2\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2, \\
&\leq \frac{4\sigma^2}{n}L^2\tau_{\max}^2\eta^2 + 4L^2\tau_{\max}\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\left\| \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\boldsymbol{w}^{s-\tau_j(s)}) \right\|_2^2 + 2\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2,
\end{aligned} \tag{S.10}$$

where the second inequality uses the fact that $\|\sum_{i=1}^{m} \boldsymbol{x}_i\|_2^2 \leq m\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2$ for vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ and Assumption 2, and the last inequality follows from Lemma S.2. Finally, plugging (S.10) back into (S.9) gives the desired result. $\square$

3 Proof of Proposition 1

Proof. We first decompose the inner product into two terms:

$$\mathbb{E}\langle \nabla F(\boldsymbol{w}^{t-1}), \boldsymbol{g}^t \rangle = \underbrace{\mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \boldsymbol{g}^t \right\rangle}_{A} + \underbrace{\mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{t-1}) - \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \boldsymbol{g}^t \right\rangle}_{B}. \tag{S.11}$$

Subsequently, we lower bound A and B, respectively. Since $\rho_i(t) \le \tau_i(t) \le \tau_{\max}$ for all $i \in [n]$, then $t - \rho_i(t) \ge t - \tau_{\max}$ for all $i \in [n]$, which implies that $\boldsymbol{\xi}_1^{t-\rho_1(t)}, \ldots, \boldsymbol{\xi}_n^{t-\rho_n(t)}$ are independent of $\boldsymbol{w}^{[t-\tau_{\max}]_+}$. Then, we have

$$\begin{aligned}
A &= \mathbb{E}\left[\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \boldsymbol{g}^t \right\rangle\right] \overset{(a)}{=} \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\xi}_1^{t-\rho_1(t)},\ldots,\boldsymbol{\xi}_n^{t-\rho_n(t)}}\left[\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \boldsymbol{g}^t \right\rangle\right]\right] \\
&= \mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \mathbb{E}_{\boldsymbol{\xi}_i^{t-\rho_i(t)}}\left[\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{w}^{t-\tau_i(t)}; \boldsymbol{\xi}_i^{t-\rho_i(t)})\right]\right\rangle \\
&\overset{(b)}{=} \mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\rangle \\
&= \underbrace{\mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1}), \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\rangle}_{A_1} + \underbrace{\mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{t-1}), \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\rangle}_{A_2},
\end{aligned} \tag{S.12}$$

where (a) uses the law of total expectation, and (b) holds due to Assumption 3. Then, we lower bound A_1 as follows:

$$\begin{aligned}
A_1 &= \mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1}), \frac{1}{n}\sum_{i=1}^{n}\left(\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-1})\right)\right\rangle \\
&\quad + \mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1}), \nabla F(\boldsymbol{w}^{t-1})\right\rangle \\
&\ge -\frac{1}{2}\mathbb{E}\|\nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-1})\right)\right\|_2^2 \\
&\quad - \mathbb{E}\|\nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{1}{4}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 \\
&= -\frac{1}{4}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{3}{2}\mathbb{E}\|\nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}) - \nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)}) - \nabla F_i(\boldsymbol{w}^{t-1})\right)\right\|_2^2,
\end{aligned} \tag{S.13}$$

where the inequality uses the fact that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq -\frac{1}{2}\|\boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{y}\|_2^2$ and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq -\|\boldsymbol{x}\|_2^2 - \frac{1}{4}\|\boldsymbol{y}\|_2^2$ for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Using the identity $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{2}\|\boldsymbol{x}\|_2^2 + \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$A_2 = \frac{1}{2}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 - \frac{1}{2}\mathbb{E}\left\|\nabla F(\boldsymbol{w}^{t-1}) - \frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2. \quad \text{(S.14)}$$

Substituting (S.13) and (S.14) back to (S.12) gives

$$\begin{aligned}
A &\geq \frac{1}{4}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{3}{2}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1}) - \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+})\|_2^2 \\
&\quad - \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla F_i(\boldsymbol{w}^{t-1}) - \nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right)\right\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 \\
&\overset{(a)}{\geq} \frac{1}{4}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - \frac{3L^2}{2}\mathbb{E}\|\boldsymbol{w}^{t} - \boldsymbol{w}^{[t-\tau_{\max}]_+}\|_2^2 \\
&\quad - \frac{L^2}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{w}^{t} - \boldsymbol{w}^{t-\tau_i(t)}\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2, \\
&\overset{(b)}{\geq} \frac{1}{4}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - 5L^2\tau_{\max}^2\eta^2\frac{\sigma^2}{n} + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 \\
&\quad - 5L^2\tau_{\max}\eta^2\sum_{s=1+[t-\tau_{\max}]_+}^{t}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2,
\end{aligned} \quad \text{(S.15)}$$

where (a) uses Assumption 2 and (b) uses Lemma S.2. To lower bound B, we observe that

$$\begin{aligned}
B &= \mathbb{E}\left\langle \nabla F(\boldsymbol{w}^{t-1}) - \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+}), \boldsymbol{g}^t \right\rangle \\
&\overset{(a)}{\geq} -\mathbb{E}\left[\|\nabla F(\boldsymbol{w}^{t-1}) - \nabla F(\boldsymbol{w}^{[t-\tau_{\max}]_+})\|_2\|\boldsymbol{g}^t\|_2\right] \\
&\overset{(b)}{\geq} -L\mathbb{E}\left[\|\boldsymbol{w}^{t} - \boldsymbol{w}^{[t-\tau_{\max}]_+}\|_2\|\boldsymbol{g}^t\|_2\right] \\
&\overset{(c)}{=} -L\mathbb{E}\left[\left\|\sum_{s=1+[t-\tau_{\max}]_+}^{t}\eta\boldsymbol{g}^s\right\|_2\|\boldsymbol{g}^t\|_2\right] \\
&\overset{(d)}{\geq} -L\mathbb{E}\left[\sum_{s=1+[t-\tau_{\max}]_+}^{t}\eta\|\boldsymbol{g}^s\|_2\|\boldsymbol{g}^t\|_2\right] \\
&\overset{(e)}{\geq} -L\eta\sum_{s=1+[t-\tau_{\max}]_+}^{t}\frac{1}{2}\left(\mathbb{E}\|\boldsymbol{g}^s\|_2^2 + \mathbb{E}\|\boldsymbol{g}^t\|_2^2\right) \\
&= -\frac{L\eta}{2}\sum_{s=1+[t-\tau_{\max}]_+}^{t}\mathbb{E}\|\boldsymbol{g}^s\|_2^2 - \frac{L\eta}{2}\tau_{\max}\mathbb{E}\|\boldsymbol{g}^t\|_2^2,
\end{aligned} \quad \text{(S.16)}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) follows from Assumption 2, (c) uses the telescoping sum

$$\boldsymbol{w}^t - \boldsymbol{w}^{[t-\tau_{\max}]_+} = \sum_{s=1+[t-\tau_{\max}]_+}^{t} \left(\boldsymbol{w}^s - \boldsymbol{w}^{s-1}\right) = \sum_{s=1+[t-\tau_{\max}]_+}^{t} \eta \boldsymbol{g}^s,$$

(d) uses the triangle inequality and (e) is due to the Young's inequality. Applying Lemma S.3 to (S.16) and simplifying, we have

$$\begin{aligned}
B \geq & -(2L\tau_{\max}\eta + 8L^3\tau_{\max}^3\eta^3)\frac{\sigma^2}{n} - 2L\eta \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 \\
& - 2L\tau_{\max}\eta\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - 4L^3\tau_{\max}^2\eta^3 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2 \\
& - 4L^3\tau_{\max}\eta^3 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \sum_{s'=1+[s-\tau_{\max}]_+}^{s} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s'-\tau_i(s')})\right\|_2^2
\end{aligned}$$

Further using the fact that $\sum_{s=1+[t-K]_+}^{t}\sum_{s'=1+[s-K]_+}^{s} a_{s'} \leq K\sum_{s=1+[t-2K]_+}^{t} a_s$ for $a_1,\ldots,a_t \geq 0$ and $K \geq 1$, we have

$$\begin{aligned}
B \geq & -(2L\tau_{\max}\eta + 8L^3\tau_{\max}^3\eta^3)\frac{\sigma^2}{n} - 2L\eta \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 \\
& - 2L\tau_{\max}\eta\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - 8L^3\tau_{\max}^2\eta^3 \sum_{s=1+[t-2\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2. \qquad \text{(S.17)}
\end{aligned}$$

Substituting (S.15) and (S.17) into (S.11) and simplifying, we have

$$\begin{aligned}
\mathbb{E}\langle\nabla F(\boldsymbol{w}^{t-1}), \boldsymbol{g}^t\rangle \geq & \left(\frac{1}{4} - 2L\tau_{\max}\eta\right)\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - 2L\eta \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 \\
& - (2L\tau_{\max}\eta + 5L^2\tau_{\max}^2\eta^2 + 8L^3\tau_{\max}^3\eta^3)\frac{\sigma^2}{n} + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 \\
& - (5L^2\tau_{\max}\eta^2 + 8L^3\tau_{\max}^2\eta^3) \sum_{s=1+[t-2\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2 \\
\geq & \frac{1}{8}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 - 2L\eta \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 - 3L\tau_{\max}\eta\frac{\sigma^2}{n} \\
& + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 - 6L^2\tau_{\max}\eta^2 \sum_{s=1+[t-2\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2,
\end{aligned}$$

where the last inequality holds because the stepsize condition $\eta \le 1/(16L\tau_{\max})$ implies that

$$\frac{1}{4} - 2L\tau_{\max}\eta \ge \frac{1}{8},$$
$$2L\tau_{\max}\eta + 5L^2\tau_{\max}^2\eta^2 + 8L^3\tau_{\max}^3\eta^3 \le 2L\tau_{\max}\eta + 6L^2\tau_{\max}^2\eta^2 \le 3L\tau_{\max}\eta,$$
$$5L^2\tau_{\max}\eta^2 + 8L^3\tau_{\max}^2\eta^3 \le 6L^2\tau_{\max}\eta^2.$$

This completes the proof. □

4 Proof of Theorem 1

Proof. Since F is L-smooth, it follows from the descent lemma that

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^t)] - \mathbb{E}[F(\boldsymbol{w}^{t-1})] &\le \mathbb{E}[\langle \nabla F(\boldsymbol{w}^{t-1}), \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle] + \frac{L}{2}\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t-1}\|_2^2 \\ &= -\eta\mathbb{E}\langle \nabla F(\boldsymbol{w}^{t-1}), \boldsymbol{g}^t\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\boldsymbol{g}^t\|_2^2. \end{aligned} \tag{S.18}$$

Substituting Lemma S.3 and Proposition 1 into (S.18), we obtain

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{w}^t)] - \mathbb{E}[F(\boldsymbol{w}^{t-1})] \\
\le& -\frac{1}{8}\eta\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 + 2L\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 + 3L\tau_{\max}\eta^2\frac{\sigma^2}{n} \\
&- \frac{1}{2}\eta\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 + 6L^2\tau_{\max}\eta^3 \sum_{s=1+[t-2\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2 \\
&+ \left(L\eta^2 + 4L^3\tau_{\max}^2\eta^4\right)\frac{\sigma^2}{n} + 4L^3\tau_{\max}\eta^4 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2 \\
&+ 2L\eta^2\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 \\
\le& -\left(\frac{1}{8}\eta - 2L\eta^2\right)\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 + 2L\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 \\
&+ (L\eta^2 + 3L\tau_{\max}\eta^2 + 4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n} - \frac{1}{2}\eta\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2 \\
&+ (6L^2\tau_{\max}\eta^3 + 4L^3\tau_{\max}\eta^4) \sum_{s=1+[t-2\tau_{\max}]_+}^{t} \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2 \\
\le& -\frac{1}{16}\eta\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2 + 2L\eta^2 \sum_{s=1+[t-\tau_{\max}]_+}^{t} \mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2 + (4L\tau_{\max}\eta^2 + 4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n}
\end{aligned}$$

$$-\frac{1}{2}\eta\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{w}^{t-\tau_i(t)})\right\|_2^2+7L^2\tau_{\max}\eta^3\sum_{s=1+[t-2\tau_{\max}]_+}^{t}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2, \tag{S.19}$$

where the last inequality holds because $\tau_{\max}\geq 1 \Rightarrow L\eta^2+3L\tau_{\max}\eta^2\leq 4L\tau_{\max}\eta$ and by requiring the following stepsize conditions:

$$\eta\leq\frac{1}{32L}\Longleftrightarrow\frac{1}{8}\eta-2L\eta^2\geq\frac{1}{16}\eta,\quad \eta\leq\frac{1}{4L}\Longrightarrow 6L^2\tau_{\max}\eta^3+4L^3\tau_{\max}\eta^4\leq 7L^2\tau_{\max}\eta^3. \tag{S.20}$$

Summing up both sides of the inequality (S.19) for $t=1,\ldots,T$ yields

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{w}^T)]-F(\boldsymbol{w}^0)\\
\leq&-\frac{1}{16}\eta\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2+2L\eta^2\sum_{t=1}^{T}\sum_{s=1+[t-\tau_{\max}]_+}^{t}\mathbb{E}\|\nabla F(\boldsymbol{w}^{s-1})\|_2^2+\sum_{t=1}^{T}(4L\tau_{\max}\eta^2+4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n}\\
&-\frac{1}{2}\eta\sum_{t=1}^{T}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\|_2^2+7L^2\tau_{\max}\eta^3\sum_{t=1}^{T}\sum_{s=1+[t-2\tau_{\max}]_+}^{t}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{s-\tau_j(s)})\right\|_2^2\\
\leq&-\frac{1}{16}\eta\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2+2L\tau_{\max}\eta^2\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2+\sum_{t=1}^{T}(4L\tau_{\max}\eta^2+4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n}\\
&-\frac{1}{2}\eta\sum_{t=1}^{T}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\|_2^2+14L^2\tau_{\max}^2\eta^3\sum_{t=1}^{T}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\|_2^2\\
=&-\left(\frac{1}{16}\eta-2L\tau_{\max}\eta^2\right)\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2+\sum_{t=1}^{T}(4L\tau_{\max}\eta^2+4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n}\\
&-\left(\frac{1}{2}\eta-14L^2\tau_{\max}^2\eta^3\right)\sum_{t=1}^{T}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\boldsymbol{w}^{t-\tau_j(t)})\right\|_2^2\\
\leq&-\frac{1}{32}\eta\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{w}^{t-1})\|_2^2+\sum_{t=1}^{T}(4L\tau_{\max}\eta^2+4L^3\tau_{\max}^2\eta^4)\frac{\sigma^2}{n},
\end{aligned} \tag{S.21}$$

where the second inequality uses the fact that $\sum_{t=1}^{T}\sum_{s=1+[t-K]_+}^{t}a_s\leq K\sum_{t=1}^{T}a_t$ for $a_1,\ldots,a_T\geq 0$ and $K\geq 1$, and the last inequality holds by requiring the following stepsize conditions:

$$\eta\leq\frac{1}{64L\tau_{\max}}\Longleftrightarrow\frac{1}{16}\eta-2L\tau_{\max}\eta^2\geq\frac{1}{32}\eta,\quad \eta\leq\frac{1}{\sqrt{28}L\tau_{\max}}\Longleftrightarrow\frac{1}{2}\eta-14L^2\tau_{\max}^2\eta^3\geq 0. \tag{S.22}$$

Note that the stepsize conditions (S.20) and (S.22) are uniformly implied by $\eta\leq\frac{1}{64L\tau_{\max}}$. Rearranging (S.21) and using Assumption 1, we obtain the convergence bound (13) in Theorem 1. □