Asynchronous SGD

December 10, 2024

Abstract

TBA.

1 Previous algorithm

1.1 Assumptions

Assumption 1. Local functions f_i are differentiable and L-smooth for some positive constant L, namely,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 2. Stochastic gradients $g_i(x) = \nabla f_i(x, \xi)$ are unbiased estimators of $\nabla f_i(x)$, *i.e.*,

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} [\nabla f_i(x, \xi)] = \nabla f_i(x), \quad \forall x \in \mathbb{R}^d,$$

and have bounded variance $\sigma^2 \geq 0$, namely,

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} \left[\|\nabla f_i(x, \xi) - \nabla f_i(x)\|^2\right] \leq \sigma^2, \quad \forall x \in \mathbb{R}^d.$$

Next, we also assume that the bounded function heterogeneity assumption holds since in general case it is not possible to derive any convergence guarantees for asynchronous algorithms.

Assumption 3. Local gradients $\nabla f_i(x)$ satisfy bounded heterogeneity condition for some $\zeta^2 \geq 0$, *i.e.*,

$$\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \zeta^2, \quad \forall x \in \mathbb{R}^d.$$

1.2 Notations

Definition 0. Corresponding delays: $\tau_t, \tilde{\tau}_t \geq 0$, then

$$\pi_t := t - \tau_t, \quad \alpha_t := t - \tilde{\tau}_t.$$

Definition 1. Let $\{\tau_t\}_{t=0}^{T-1}$ be the delays of all applied gradients.
The average and maximum delays are defined as follows:

$$\tau_{\text{avg}} := \frac{1}{|\mathcal{A}_{T+1}|} \left(\sum_{t=0}^{T-1} \tau_t + \sum_{(i,j) \in \mathcal{A}_{T+1} \setminus \mathcal{R}_T} T - j \right), \quad \tau_{\max} := \max \left\{ \max_{0 \leq t < T} \tau_t, \max_{(i,j) \in \mathcal{A}_{T+1} \setminus \mathcal{R}_T} T - j \right\}.$$

Definition 2. The maximum number of active jobs or concurrency is defined as

$$\tau_C := \max_{0 \leq t \leq T} |\mathcal{A}_{t+1} \setminus \mathcal{R}_t| .$$

Definition 3.

$$\widetilde{x}_0 = x_0, \ \widetilde{x}_{t+1} = \begin{cases} \widetilde{x}_t - \gamma \nabla f(x_t) & \text{if } t+1 \neq 0 \mod \tau, \\ x_{t+1} & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

where $\tau = \Theta(\frac{1}{L\gamma})$.

1.3 Pure Asynchronous SGD

1.3.1 Algorithm

Algorithm 1 Pure Asynchronous SGD

Input: initial point x_0, stepsize γ, set of assigned jobs $\mathcal{A}_0 = \varnothing,\ \mathcal{A}_1 = \{(i,0)\colon i \in [n]\}$, set of received jobs $\mathcal{R}_0 = \varnothing$

1: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
2: once worker i_t finishes a job $(i_t, \pi_t) \in \mathcal{A}_{t+1}$(computing $g_{i_t}(x_{\pi_t})$), it sends $g_{i_t}(x_{\pi_t})$ to the server
3: server updates the current model $x_{t+1} = x_t - \gamma g_{i_t}(x_{\pi_t})$ and the set $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{(i_t, \pi_t)\}$
4: server assigns worker i_t to compute a gradient $g_{i_t}(x_{t+1})$
5: server updates the set $\mathcal{A}_{t+2} = \mathcal{A}_{t+1} \cup \{(i_t, t+1)\}$
6: **end for**

1.3.2 Lemmas

For $r(t) \le m < r(t) + \tau$ $(r(t) = k\tau)$, Denote

$$A := \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right],$$

$$B := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right],$$

Virtual and real iterates:

$$x_t = x_{r(t)} - \gamma \sum_{j=r(t)}^{t-1} g_{i_j}(x_{\pi_j})$$

$$\widetilde{x}_t = x_{r(t)} - \gamma \sum_{j=r(t)}^{t-1} \nabla f(x_j)$$

$$\Delta_t^m := \sum_{j=r(t)}^{m} \left(\nabla f(x_j) - g_{i_j}(x_{\pi_j})\right)$$

Then we have some useful lemmas below:

Lemma 1.

$$\mathbb{E}\left[\|\Delta_t^m\|^2\right] \leq \mathbb{E}\left[\left\|\sum_{j=r(t)}^{m} \nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_j)\right\|^2\right] + \tau\sigma^2$$
$$\leq 2\tau^2\zeta^2 + 2L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] + \tau\sigma^2$$

(Here's sth. wrong. I didn't consider about $r(t)$. However, it doesn't affect the final result.)

Lemma 2.

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - \widetilde{x}_t\|^2\right] =&\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{j=r(t)}^{t-1} g_{i_j}(x_{\pi_j}) - \nabla f(x_j)\right\|^2\right] \\
=&\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|\Delta_t^{t-1}\|^2\right] \\
\leq& 2\gamma^2\tau^2\zeta^2 T + 2\gamma^2 L^2 \tau \sum_{t=0}^{T-1} \sum_{j=r(t)}^{t-1} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] + \gamma^2\tau\sigma^2 T \\
\leq& 2\gamma^2\tau^2\zeta^2 T + 2\gamma^2 L^2 \tau^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right] + \gamma^2\tau\sigma^2 T
\end{aligned}$$

Lemma 3. If $20\gamma L\sqrt{\tau_{\max}\tau_C} \leq 1$, recall that $\pi_t = t - \tau_t$,

$$\begin{aligned}
&\mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right] \\
=&\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} g_{i_j}(x_{\pi_j})\right\|^2\right] \\
\leq&\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f_{i_j}(x_{\pi_j})\right\|^2\right] + \gamma^2(t-\pi_t)\sigma^2 \\
\leq&3\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} (\nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_{\pi_j}))\right\|^2\right] + 3\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} (\nabla f(x_{\pi_j}) - \nabla f(x_j))\right\|^2\right] + 3\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f(x_j)\right\|^2\right] + \tau_t\gamma^2\sigma \\
\leq&3\gamma^2 L^2 \tau_t \sum_{j=\pi_t}^{t-1} \mathbb{E}\left[\|x_{\pi_j} - x_j\|^2\right] + 3\gamma^2 \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} (\nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_{\pi_j}))\right\|^2\right] + 3\gamma^2\tau_t \sum_{j=\pi_t}^{t-1} \|\nabla f(x_j)\|^2 + \tau_t\gamma^2\sigma^2.
\end{aligned}$$

Then, we sum it up from 0 to $T-1$. By $20\gamma L\sqrt{\tau_{\max}\tau_C} \leq 1$, we can get

$$
\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[\|x_{\pi_t} - x_t\|^2\right] \leq & 3\gamma^2 L^2 \tau_{\max} \sum_{t=0}^{T-1}\sum_{j=\pi_t}^{t-1} \mathbb{E}\left[\left\|x_{\pi_j} - x_j\right\|^2\right] + 3\gamma^2 \tau_{\max} \sum_{t=0}^{T-1}\sum_{j=\pi_t}^{t-1} \mathbb{E}\left[\|\nabla f(x_j)\|^2\right] \\
& + 3\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_{\pi_j})\right\|^2\right] + \tau_{\text{avg}} T\gamma^2\sigma^2 \\
\leq & 3\gamma^2 L^2 \tau_{\max}\tau_C \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_{\pi_t} - x_t\|^2\right] + 3\gamma^2 \tau_{\max}\tau_C \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
& + 3\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f_{i_j}(x_j) - \nabla f(x_j)\right\|^2\right] + \tau_{\text{avg}} T\gamma^2\sigma^2 \\
\leq & \frac{3}{400} \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_{\pi_t} - x_t\|^2\right] + \frac{3}{400L^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
& + 3\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_{\pi_j})\right\|^2\right] + \tau_{\text{avg}} T\gamma^2\sigma^2,
\end{aligned}
$$

Let τ_{sum}^t represent the sum of the delays of all tasks at the end of round $t-1$, and τ_C^t represent the maximum delay of the task active at the end of round $t-1$.

Then $\tau_{sum}^t \leq \tau_{sum}^{t-1} + \tau_C^t \Rightarrow \tau_{sum} \leq \sum \tau_C^t \Rightarrow \tau_{avg} \leq 2\tau_C$.

Thus,

$$
\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right] \leq & \frac{1}{132L^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{2\tau_{\text{avg}}}{20L\sqrt{\tau_{\max}\tau_C}} T\gamma\sigma^2 \\
& + \frac{100\gamma^2}{33} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{j=\pi_t}^{t-1} \nabla f_{i_j}(x_{\pi_j}) - \nabla f(x_{\pi_j})\right\|^2\right] \\
\leq & \frac{1}{132L^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\gamma}{5L} T\sigma^2 + \frac{\zeta^2 T}{132L^2}.
\end{aligned}
$$

Lemma 4. If $20\gamma L\tau \leq 1$ and $20\gamma L\sqrt{\tau_{\max}\tau_C} \leq 1$, by Lemma 2 and Lemma 3,

$$
\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - \widetilde{x}_t\|^2\right] &\leq 2\gamma^2\tau^2\zeta^2 T + 2\gamma^2 L^2\tau^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right] + \gamma^2\tau\sigma^2 T \\
&\leq \frac{\zeta^2 T}{200L^2} + \frac{1}{200} A + \frac{\gamma}{L} T\sigma^2 \\
&\leq \frac{\zeta^2 T}{200L^2} + \frac{1}{200}\left(\frac{1}{132L^2} B + \frac{\zeta^2 T}{132L^2} + \frac{\gamma T\sigma^2}{5L}\right) + \frac{\gamma}{L} T\sigma^2 \\
&\leq \frac{\zeta^2 T}{100L^2} + \frac{1}{20000L^2} B + \frac{2\gamma}{L} T\sigma^2
\end{aligned}
$$

1.3.3 Analysis

Proposition 1. Let Assumptions 1,2 and 3 hold. Let the stepsize γ satisfy inequalities

$$20L\gamma\sqrt{\tau_{\max}\tau_C} \le 1, \quad 6L\gamma \le 1$$

Let $\tau = \lfloor \frac{1}{20L\gamma} \rfloor$. Then the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \le \mathcal{O}\left(\frac{F_0}{\gamma T} + L\gamma\sigma^2 + \zeta^2\right),$$

where $\hat{x}_t$ is chosen uniformly at random from $\{x_0, \ldots, x_{T-1}\}$ and $F_0 := f(x_0) - f^*$.

Moreover, if we tune the stepsize, then the iterates of pure asynchronous SGD satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \le \mathcal{O}\left(\frac{LF_0\sqrt{\tau_{\max}\tau_C}}{T} + \left(\frac{LF_0\sigma^2}{T}\right)^{1/2} + \zeta^2\right)$$

Proof. First, we consider a descent inequality for the virtual iterates $\tilde{x}_t$:

$$\widetilde{x}_0 = x_0, \ \widetilde{x}_{t+1} = \begin{cases} \widetilde{x}_t - \gamma\nabla f(x_t) & \text{if } t+1 \ne 0 \mod \tau, \\ x_{t+1} & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

Iterations without restart($t+1 \ne 0 \mod \tau$):

$$\begin{aligned}
\mathbb{E}\left[f(\widetilde{x}_{t+1})\right] &\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \gamma\mathbb{E}\left[\langle\nabla f(\widetilde{x}_t), \nabla f(x_t)\rangle\right] + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
&= \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t) - \nabla f(x_t)\|^2\right] + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
&\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right] + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
&\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right].
\end{aligned}$$

Iterations with restart($t+1 = 0 \mod \tau$):

$$\begin{aligned}
\widetilde{x}_{t+1} = x_{t+1} &= x_t - \gamma g_{i_t}(x_{\pi_t}) \\
&= \widetilde{x}_t + (x_t - \widetilde{x}_t) - \gamma\nabla f(x_t) + (\gamma\nabla f(x_t) - \gamma g_{i_t}(x_{\pi_t})) \\
&= \widetilde{x}_t - \gamma\nabla f(x_t) + \gamma \underbrace{\sum_{j=r(t)}^{t} \nabla f(x_j) - g_{i_j}(x_{\pi_j})}_{=\Delta_t^t}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[f(\widetilde{x}_{t+1})\right] &\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \gamma\mathbb{E}\left[\langle\nabla f(\widetilde{x}_t), \nabla f(x_t) - \Delta_t^t\rangle\right] + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_t) - \Delta_t^t\|^2\right] \\
&\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \gamma\mathbb{E}\left[\langle\nabla f(\widetilde{x}_t), \nabla f(x_t)\rangle\right] + \gamma\mathbb{E}\left[\langle\nabla f(\widetilde{x}_t), \Delta_t^t\rangle\right] + L\gamma^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right] \\
&\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t) - \nabla f(x_t)\|^2\right] \\
&+ \frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] + 40L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right] + L\gamma^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right] \\
&\le \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right] + \frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] \\
&+ 41L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right]
\end{aligned}$$

Thus, let

$$\xi_t = \begin{cases} 1, & \text{if } t+1 \neq 0 \mod \tau, \\ 0, & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

Then,

$$\begin{aligned} \mathbb{E}\left[f(\widetilde{x}_{t+1})\right] \leq \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right] \\ + \left(\frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] + 41L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right]\right)\xi_t, \quad \forall t \geq 0. \end{aligned}$$

Below we estimate the two terms associated with ξ_t.

First term: the gradient at moment t is bounded by the previous τ round:

$$\begin{aligned} \frac{1}{L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] &= \frac{1}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] \\ &\leq \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t) - \nabla f(\widetilde{x}_{t-j})\|^2\right] + \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \\ &\leq \frac{2L}{\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\widetilde{x}_t - \widetilde{x}_{t-j}\|^2\right] + \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \\ &\leq \frac{2L\gamma^2}{\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\sum_{l=t-j}^{t-1}\nabla f(x_l)\right\|^2\right] + \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \\ &\leq 2L\gamma^2\sum_{j=0}^{\tau-1}\sum_{l=t-j}^{t-1}\mathbb{E}\left[\|\nabla f(x_l)\|^2\right] + \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \\ &\leq 2L\gamma^2\tau\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(x_{t-j})\|^2\right] + \frac{2}{L\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \\ &\leq \frac{\gamma}{10}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(x_{t-j})\|^2\right] + 80\gamma\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_{t-j})\|^2\right] \end{aligned}$$

By $\frac{1}{40} \leq L\gamma\tau \leq \frac{1}{20}(\tau = \lfloor\frac{1}{20L\gamma}\rfloor)$, we can get: (Just add up those terms $t = k\tau$ here.)

$$\sum_{t=0}^{T-1}\frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right]\xi_t \leq \frac{\gamma}{1600}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right]$$

Second term: By Lemma 1 and Lemma 3,

$$
\begin{aligned}
L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t^t\|^2 \xi_t \leq & 2L\gamma^2\tau\zeta^2 T + 2\gamma^2 L^3 \tau \sum_{t=0}^{T-1} \sum_{j=r(t)}^{t-1} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] \xi_t + L\gamma^2\sigma^2 T \\
\leq & 2L\gamma^2\tau\zeta^2 T + 2\gamma^2 L^3 \tau A + L\gamma^2\sigma^2 T \\
\leq & 2L\gamma^2\tau\zeta^2 T + 2\gamma^2 L^3 \tau \left(\frac{1}{132L^2} B + \frac{\zeta^2 T}{132L^2} + \frac{\gamma T \sigma^2}{5L}\right) + L\gamma^2\sigma^2 T \\
\leq & 3L\gamma^2\tau\zeta^2 T + \frac{1}{66}\gamma^2 L\tau B + 2L\gamma^2\sigma^2 T
\end{aligned}
$$

Plug the two terms back and sum it up from 0 to $T-1$, and by Lemma 4,

$$
\begin{aligned}
\mathbb{E}\left[f(\widetilde{x}_T) - f(\widetilde{x}_0)\right] \leq & -\frac{\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right] \\
& + \frac{1}{160L}\sum_{t=0}^{T-1} \xi_t \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] + 41L\gamma^2 \sum_{t=0}^{T-1} \xi_t \mathbb{E}\left[\|\Delta_t^t\|^2\right] \\
\leq & -\frac{\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3} B \\
& + \frac{L^2\gamma}{2}\left(\frac{\zeta^2 T}{100L^2} + \frac{1}{20000L^2} B + \frac{2\gamma}{L} T\sigma^2\right) \\
& + \frac{\gamma}{1600} B + \frac{\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] \\
& + 124L\gamma^2\tau\zeta^2 T + \gamma^2 L\tau B + 82L\gamma^2\sigma^2 T \\
\leq & -\frac{\gamma}{4} B + 7\gamma T\zeta^2 + 83L\gamma^2\sigma^2 T
\end{aligned}
$$

Let $F_0 := f(x_0) - f^*$, the final rate

$$
\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{F_0}{\gamma T} + L\gamma\sigma^2 + \zeta^2\right).
$$

Since $\gamma \leq \dfrac{1}{L\sqrt{\tau_{\max}\tau_C}}$,

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] & \leq \mathcal{O}\left(\frac{F_0}{T}\sqrt{L\tau_{\max}\tau_C} + L\sigma^2\left(\frac{F_0}{L\sigma^2 T}\right)^{1/2} + \zeta^2\right) \\
& = \mathcal{O}\left(\frac{LF_0\sqrt{\tau_{\max}\tau_C}}{T} + \left(\frac{LF_0\sigma^2}{T}\right)^{1/2} + \zeta^2\right)
\end{aligned}
$$

2 Our algorithm

2.1 Pure

Algorithm 2 AlgoA

Input: initial point x_0, $\{h_i^0\}_{i=1}^n$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$, stepsize γ, α, set of assigned jobs $\mathcal{A}_0 = \varnothing$, $\mathcal{A}_1 = \{(i, 0) : i \in [n]\}$, set of received jobs $\mathcal{R}_0 = \varnothing$,

for $t = 0, 1, 2, \cdots, T-1$ **do**
- worker i_t finishes a job $(i_t, \pi_t) \in \mathcal{A}_{t+1}$(compute $\widetilde{g}_{i_t}(x_{\pi_t})$ and send $\widehat{\Delta}_{i_t}^t$ to server)
- $\widehat{\Delta}_{i_t}^t = \mathcal{C}_{i_t}^t(\widetilde{g}_{i_t}(x_{\pi_t}) - h_{i_t}^t)$
- $h_{i_t}^{t+1} = h_{i_t}^t + \alpha \widehat{\Delta}_{i_t}^t$
- server updates the current model $x^{t+1} = x^t - \gamma(h^t + \widehat{\Delta}_{i_t}^t)$ and the set $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{(i_t, \pi_t)\}$
- // $g_{i_t}^t = h^t + \widehat{\Delta}_{i_t}^t$
- // $\mathbb{E}[g_{i_t}^t] = h^t + \nabla f_{i_t}(x_{\pi_t}) - h_{i_t}^t$ $\qquad h^t = \frac{1}{n}\sum_i h_i^t \qquad \nabla f(x) = \frac{1}{n}\sum_i \nabla f_i(x)$
- $h^{t+1} = h^t + \frac{\alpha}{n}\widehat{\Delta}_{i_t}^t$
- server assigns worker i_t to compute a gradient $\widetilde{g}_{i_t}(x_{t+1})$
- server updates the set $\mathcal{A}_{t+2} = \mathcal{A}_{t+1} \cup \{(i_t, t+1)\}$

end for

Zhize: add our analysis

We consider a descent inequality for the virtual iterates $\tilde{x}_t$:

$$\widetilde{x}_0 = x_0, \ \widetilde{x}_{t+1} = \begin{cases} \widetilde{x}_t - \gamma \nabla f(x_t) & \text{if } t+1 \neq 0 \mod \tau, \\ x_{t+1} & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

For $r(t) \leq m < r(t) + \tau$ $(r(t) = k\tau)$, Denote

$$A := \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right],$$

$$B := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right],$$

$$[\sigma_t^m]^2 := \sum_{j=r(t)}^{m} \mathbb{E}\left[\left\|\nabla f_{i_j}(x_{\pi_j}) - h_{i_j}^j\right\|^2\right]$$

Virtual and real iterates:

$$x_t = x_{r(t)} - \gamma \sum_{j=r(t)}^{t-1} g_{i_j}^j$$

$$\widetilde{x}_t = x_{r(t)} - \gamma \sum_{j=r(t)}^{t-1} \nabla f(x_j)$$

$$\Delta_t^m := \sum_{j=r(t)}^{m} (\nabla f(x_j) - g_{i_j}^j)$$

$$\mathbb{E}_t[g_{i_t}^t] = \nabla f_{i_t}(x_{\pi_t}) + h^t - h_{i_t}^t$$

$$\begin{aligned}\mathbb{E}_t\left[\|\nabla f_{i_{t+1}}(x_{\pi_{t+1}}) - h_{i_{t+1}}^{t+1}\|^2\right] \leq &\left[1 - 2\alpha + \frac{(1-\alpha)^2}{\beta} + \alpha^2(1+\omega)\right]\|\nabla f_{i_t}(x_{\pi_t}) - h_{i_t}^t\|^2 \\ &+ (1+\beta)\|\nabla f_{i_{t+1}}(x_{\pi_{t+1}}) - \nabla f_{i_t}(x_{\pi_t}))\|^2 + \alpha^2(1+\omega)\sigma^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\|\Delta_t^m\|^2\right] \leq &4\mathbb{E}\left[\phi_t^m(x_{r(t)})\right] + 4L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] + 8L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{r(t)}\|^2\right] \\ &+\omega \sum_{j=r(t)}^{m} \mathbb{E}\|\nabla f_{i_j}(x_{\pi_j}) - h_{i_j}\|^2\end{aligned}$$

$$\begin{aligned}\sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{r(t)}\|^2\right] &= \gamma^2 \sum_{j=r(t)}^{m} \mathbb{E}\left[\left\|\sum_{l=r(t)}^{j-1} g_{i_l}^l\right\|^2\right] \\ &\leq 2\gamma^2 \sum_{j=r(t)}^{m} \mathbb{E}\left[\left\|\sum_{l=r(t)}^{j-1} (g_{i_l}^l - \nabla f(x_l))\right\|^2\right] + 2\gamma^2 \sum_{j=r(t)}^{m} \mathbb{E}\left[\left\|\sum_{l=r(t)}^{j-1} \nabla f(x_l)\right\|^2\right]\end{aligned}$$

2.2 Random

Algorithm 3 AlgoA

Input: initial point x^0, $\{h_i^0\}_{i=1}^n$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$, stepsize γ, α, set of assigned jobs $\mathcal{A}^0 = \varnothing$, $\mathcal{A}^1 = \{(i,0) \colon i \in [n]\}$, set of received jobs $\mathcal{R}^0 = \varnothing$,

for $t = 0, 1, 2, \cdots, T-1$ **do**
  worker i^t finishes a job $(i^t, \pi^t) \in \mathcal{A}^{t+1}$(compute $g_{i^t}^t(x^{\pi^t})$ and send $\widehat{\Delta}_{i^t}^t$ to server)
  $\widehat{\Delta}_{i^t}^t = \mathcal{C}_{i^t}^t(g_{i^t}^t(x^{\pi^t}) - h_{i^t}^t)$
  $h_{i^t}^{t+1} = h_{i^t}^t + \alpha \widehat{\Delta}_{i^t}^t$
  server updates the current model $x^{t+1} = x^t - \gamma(h^t + \widehat{\Delta}_{i^t}^t)$ and the set $\mathcal{R}^{t+1} = \mathcal{R}^t \cup \{(i^t, \pi^t)\}$
  // $\widetilde{g}_{i^t}^t(x^{\pi^t}) = h^t + \widehat{\Delta}_{i^t}^t$
  // $h^t = \frac{1}{n}\sum_i h_i^t$ $\nabla f(x) = \frac{1}{n}\sum_i \nabla f_i(x)$
  $h^{t+1} = h^t + \frac{\alpha}{n}\widehat{\Delta}_{i^t}^t$
  server assigns worker $k^{t+1} \sim Uni[1, \cdots, n]$ to compute a gradient $g_{k^{t+1}}(x^{t+1})$
  server updates the set $\mathcal{A}^{t+2} = \mathcal{A}^{t+1} \cup \{(k^{t+1}, t+1)\}$
end for

Zhize: add our analysis

We consider a descent inequality for the virtual iterates $\tilde{y}^t$:

$$\widetilde{y}^1 = y^1, \;\; \widetilde{y}^{t+1} = \begin{cases} \widetilde{y}^t - \gamma(h^{l^t} + \nabla f(x^t) - h_{k^t}^{l^t}) & \text{if } t \neq 0 \mod \tau, \\ y^{t+1} & \text{if } t = 0 \mod \tau. \end{cases}$$

For $r(t) \le m < r(t) + \tau$ $(r(t) = k\tau)$, Denote

$$A := \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right],$$

$$B := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right],$$

$$[\sigma_t^m]^2 := \sum_{j=r(t)}^{m} \mathbb{E}\left[\left\|\nabla f_{i_j}(x_{\pi_j}) - h_{i_j}^j\right\|^2\right]$$

Virtual and real iterates $(l := \pi^{-1})$:

$$x^t = x^{r(t)} - \gamma \sum_{j=r(t)}^{t-1} \widetilde{g}_{i_j}^j(x^{\pi^j})$$

$$y^t = y^{r(t)} - \gamma \sum_{j=r(t)}^{t-1} \widetilde{g}_{k^j}^{l^j}(x^j)$$

$$\widetilde{y}^t = y^{r(t)} - \gamma \sum_{j=r(t)}^{t-1} \nabla f(x^j)$$

$$\Delta_t^m = \sum_{j=r(t)}^{m} \nabla f(x^j) - h_{k^j}^{l^j} - \mathcal{C}_{k^j}^{l^j}(g_{k^j}^{l^j}(x^j) - h_{k^j}^{l^j})$$

Lemma D.1'

$$x^t - y^t = \gamma \sum_{(i,j)\in\mathcal{A}^t\setminus\mathcal{R}^t} \widetilde{g}_i^{l^j}(x^j)$$

Lemma D.2'

$$\mathbb{E}\left[\|y^t - x^t\|^2\right] = \gamma^2 \mathbb{E}\left[\left\| \sum_{(i,j)\in\mathcal{A}^t\setminus\mathcal{R}^t} \widetilde{g}_i^{l^j}(x^j)\right\|^2\right]$$

Lemma D.3'

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta_t^m\|^2\right] \\
=&\mathbb{E}\left[\left\|\sum_{j=r(t)}^{m} \nabla f(x^j) - h_{k^j}^{l^j} - \mathcal{C}_{k^j}^{l^j}(g_{k^j}^{l^j}(x^j) - h_{k^j}^{l^j})\right\|^2\right] \\
\leq&
\end{aligned}$$

2.3 Pure (SAGA)

Algorithm 4 AlgoA

Input: initial point $\{x^0 = w_i^0\}_{i=1}^n$, stepsize γ, set of assigned jobs $\mathcal{A}^0 = \varnothing,\ \mathcal{A}^1 = \{(i,0)\colon i \in [n]\}$, set of received jobs $\mathcal{R}^0 = \varnothing$,

1: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
2: worker i^t finishes a job $(i^t, \pi^t) \in \mathcal{A}^{t+1}$(compute $\nabla f_{i^t}(x^{\pi^t})$ and send it to server)
3: $\widetilde{g}_{i^t}(x^{\pi^t}) = \nabla f_{i^t}(x^{\pi^t}) - \nabla f_{i^t}(w_{i^t}^t) + \frac{1}{n}\sum_{p=1}^n \nabla f_p(w_p^t)$
4: server updates the current model $x^{t+1} = x^t - \gamma \widetilde{g}_{i^t}(x^{\pi^t})$ and the set $\mathcal{R}^{t+1} = \mathcal{R}^t \cup \{(i^t, \pi^t)\}$
5: server assigns worker i^t to compute $\nabla f_{i^t}(x^{t+1})$
6: server updates the set $\mathcal{A}^{t+2} = \mathcal{A}^{t+1} \cup \{(k^{t+1}, t+1)\}$
7: $w_i^{t+1} = \begin{cases} x^{\pi^t}, & i = i^t \\ w_i^t, & i \neq i^t \end{cases}$
8: **end for**

Zhize: add our analysis

Virtual and real iterates ($l := \pi^{-1}$, where $\pi(\cdot) \neq 0$):

$$x^{t+1} = x^t - \gamma \widetilde{g}_{i^t}(x^{\pi^t})$$

$$y_{t+1} = y_t - \gamma \sum_{(i,j) \in \mathcal{A}_{t+1} \setminus \mathcal{A}_t} \widetilde{g}_i(x^j) \overset{t \geq 0}{=} y_t - \gamma \widetilde{g}_{k^t}(x^t)$$

$$\widetilde{g}_i(x^0) = \nabla f_i(x^0) - \nabla f_i(x^0) + \frac{1}{n}\sum_{p=1}^n \nabla f_p(x^0) = \nabla f(x^0)$$

(used to complete the definition during the proof below)

$$\begin{aligned} &\mathbb{E}\left[\|y^t - x^t\|^2\right] \\ =&\gamma^2 \mathbb{E}\left[\left\|\sum_{(i,j)\in\mathcal{A}^t \setminus \mathcal{R}^t} \widetilde{g}_i(x^j)\right\|^2\right] \end{aligned}$$

2.4 Pure (revised SAGA)

Algorithm 5 AlgoA

Input: initial point $\{x^0 = w_i^0\}_{i=1}^n$, stepsize γ, set of assigned jobs $\mathcal{A}^0 = \varnothing$, $\mathcal{A}^1 = \{(i,0)\colon i \in [n]\}$, set of received jobs $\mathcal{R}^0 = \varnothing$,

for $t = 0, 1, 2, \cdots, T-1$ **do**

worker i^t finishes a job $(i^t, \pi^t) \in \mathcal{A}^{t+1}$(compute $\nabla f_{i^t}(x^{\pi^t})$ and send it to server)

$w_i^t = \begin{cases} x^{\pi^t}, & i = i^t \\ w_i^{t-1}, & i \neq i^t \end{cases}$

$g^t = \frac{1}{n}\sum_{p=1}^n \nabla f_p(w_p^t)$

server updates the current model $x^{t+1} = x^t - \gamma g^t$ and the set $\mathcal{R}^{t+1} = \mathcal{R}^t \cup \{(i^t, \pi^t)\}$

server assigns worker i^t to compute $\nabla f_{i^t}(x^{t+1})$

server updates the set $\mathcal{A}^{t+2} = \mathcal{A}^{t+1} \cup \{(k^{t+1}, t+1)\}$

end for

Zhize: add our analysis

$$A^t =: \left\|x^{t+1} - x^t\right\|^2 \tag{2.1}$$

$$\widetilde{A}^t =: \left\|x^{\pi^t} - x^t\right\|^2 \tag{2.2}$$

$$B^t =: \frac{1}{n}\sum_{p=1}^n \left\|w_p^t - x^t\right\|^2 \tag{2.3}$$

$$R^t =: f(x^t) + cB^t, \quad c = n\gamma L^2 \tag{2.4}$$

$$F^0 =: f(x^0) \tag{2.5}$$

Lemma 1.

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2}\|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) A^t + \frac{\gamma}{2}\|g^t - \nabla f(x^t)\|^2 \tag{2.6}$$

Lemma 2.

$$\mathbb{E}\left[\left\|g^t - \nabla f(x^t)\right\|^2\right] \leq L^2 \mathbb{E}\left[B^t\right] \tag{2.7}$$

Proof.

$$\mathbb{E}\left[\left\|g^t - \nabla f(x^t)\right\|\right]^2 = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{p=1}^{n}\nabla f_p(w_p^t) - \nabla f(x^t)\right\|^2\right] \tag{2.8}$$

$$\leq \frac{1}{n}\sum_{p=1}^{n}\mathbb{E}\left[\left\|\nabla f_p(w_p^t) - \nabla f(x^t)\right\|^2\right] \tag{2.9}$$

$$\leq \frac{L^2}{n}\sum_{p=1}^{n}\mathbb{E}\left[\left\|w_p^t - x^t\right\|^2\right] \tag{2.10}$$

□

Lemma 3. $\forall \beta > 0$,

$$\mathbb{E}\left[B^{t+1}\right] \leq \left(1 - \frac{1}{n}\right)(1+\beta)\,\mathbb{E}\left[B^t\right] + \frac{1}{n}\mathbb{E}\left[\widetilde{A}^{t+1}\right] + \left(1 - \frac{1}{n}\right)\left(1 + \frac{1}{\beta}\right)\mathbb{E}\left[A^t\right] \tag{2.11}$$

Proof.

$$\mathbb{E}\left[B^{t+1}\right] = \frac{1}{n}\sum_{p=1}^{n}\mathbb{E}\left[\left\|w_p^{t+1} - x^{t+1}\right\|^2\right] \tag{2.12}$$

$$= \mathbb{E}\left[\frac{1}{n}\widetilde{A}^{t+1}\right] + \left(1 - \frac{1}{n}\right)\frac{1}{n}\sum_{p=1}^{n}\mathbb{E}\left[\left\|w_p^t - x^{t+1}\right\|^2\right] \tag{2.13}$$

$$\leq \mathbb{E}\left[\frac{1}{n}\widetilde{A}^{t+1}\right] + \left(1 - \frac{1}{n}\right)\frac{1}{n}\sum_{p=1}^{n}\mathbb{E}\left[\left(1 + \frac{1}{\beta}\right)A^t + (1+\beta)\left\|w_p^t - x^{t+1}\right\|^2\right] \tag{2.14}$$

$$= \left(1 - \frac{1}{n}\right)(1+\beta)\,\mathbb{E}\left[B^t\right] + \left(1 - \frac{1}{n}\right)\left(1 + \frac{1}{\beta}\right)\mathbb{E}\left[A^t\right] + \frac{1}{n}\mathbb{E}\left[\widetilde{A}^{t+1}\right] \tag{2.15}$$

□

Lemma 4.

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{A}^{t+1}\right] \leq \tau_{\max}\tau_C\sum_{t=0}^{T-1}\mathbb{E}\left[A^t\right] \tag{2.16}$$

Proof.

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{A}^{t+1}\right] \leq \sum_{t=0}^{T-1}\tau_{\max}\sum_{l=\pi^{t+1}}^{t}\mathbb{E}\left[\left\|x^{l+1} - x^l\right\|^2\right] \tag{2.17}$$

$$\leq \tau_{\max}\tau_C\sum_{t=0}^{T-1}\mathbb{E}\left[A^t\right] \tag{2.18}$$

□

Lemma 5. *If* $\gamma \leq \dfrac{1}{4L\sqrt{\tau_{\max}\tau_C + n^2}}$,

$$\mathbb{E}\left[R^T\right] \leq \mathbb{E}\left[R^0\right] - \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \tag{2.19}$$

Proof. By Lemma 1 & Lemma 2 & Lemma 3, and let $\beta = \dfrac{1}{2n}$, we have

$$\mathbb{E}\left[R^{t+1}\right] = \mathbb{E}\left[f(x^{t+1}) + cB^{t+1}\right] \tag{2.20}$$

$$\leq \mathbb{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|\right]^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[A^t\right] + \frac{\gamma L^2}{2}\mathbb{E}\left[B^t\right] + \mathbb{E}\left[cB^{t+1}\right] \tag{2.21}$$

$$\leq \mathbb{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|\right]^2 + \left(\left(1 - \frac{1}{n}\right)(1+\beta) + \frac{\gamma L^2}{2}\right)\mathbb{E}\left[B^t\right] \tag{2.22}$$

$$+ \left(\left(1 - \frac{1}{n}\right)\left(1 + \frac{1}{\beta}\right)c - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\right)\mathbb{E}\left[A^t\right] + \frac{c}{n}\mathbb{E}\left[\widetilde{A}^{t+1}\right] \tag{2.23}$$

$$\leq \mathbb{E}\left[R^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|\right]^2 \tag{2.24}$$

$$+ \left(2n^2\gamma L^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\right)\mathbb{E}\left[A^t\right] + \gamma L^2\mathbb{E}\left[\widetilde{A}^{t+1}\right] \tag{2.25}$$

By adding summation & lemma 4, we have

$$\mathbb{E}\left[R^T\right] \leq \mathbb{E}\left[R^0\right] - \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \tag{2.26}$$

$$+ \left(2n^2\gamma L^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\right)\sum_{t=0}^{T-1}\mathbb{E}\left[A^t\right] + \gamma L^2\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{A}^{t+1}\right] \tag{2.27}$$

$$\leq \mathbb{E}\left[R^0\right] - \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \tag{2.28}$$

$$+ \left(2n^2\gamma L^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) + \gamma L^2\tau_{\max}\tau_C\right)\sum_{t=0}^{T-1}\mathbb{E}\left[A^t\right] \tag{2.29}$$

$$\leq \mathbb{E}\left[R^0\right] - \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \tag{2.30}$$

□

Theorem 1. *If* $\gamma \leq \dfrac{1}{4L\sqrt{\tau_{\max}\tau_C + n^2}}$,

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x})\right\|^2\right] \leq \mathcal{O}\left(\frac{F^0 L\sqrt{\tau_{\max}\tau_C + n^2}}{T}\right) \tag{2.31}$$

where $\widehat{x}$ *randomly chosen from* $\{x^t\}_{t=0}^{T-1}$ *with probability* $\dfrac{1}{T}$ *for* x^t.

Proof. By Lemma 5,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] &\leq \frac{2F^0}{\gamma T} && (2.32)\\
&\leq \frac{8F^0 L\sqrt{\tau_{\max}\tau_C + n^2}}{T} && (2.33)\\
&\leq \mathcal{O}\left(\frac{F^0 L\sqrt{\tau_{\max}\tau_C + n^2}}{T}\right) && (2.34)
\end{aligned}
$$

□