Appendix for Paper "Faster Derivative-Free Stochastic Algorithm for Shared Memory Machines"

Notations. To make the paper easier to follow, we give the following notations.

- x denotes the vector x in the shared memory. If reading the vector x from the shared memory to the local memory, which is denoted as $\widehat{x}$.
- e_j is the N-dimensional zero vector except that the j-th coordinate is 1.
- $\nabla_j f(x)$ is the j-th coordinate of the gradient $\nabla f(x)$.
- $\{\mu_j\}_{j=1\dots,N}$ are the approximate parameters.
- $\{\gamma_t\}_{t=0,\dots,m-1}$ are the decreasing steplengths in AsySZO, and γ is the steplength in AsySZO+.
- T, m and S are the total number of iterations, the number of iterations in the inner loop, and the number of iterations in the outer loop, respectively.
- Y is the size of the coordinate set J.
- L and $\widetilde{L}$ are the Lipschitz constants for $\nabla f_i(x)$ and $\sum_{j=1}^{N} \nabla_j f_i^j(x)$ respectively.

1. Convergence Analysis

In this section, we prove the convergence rate of AsySZO+ (Theorem 2 and Corollary 1). Specifically, we improve the convergence rate of asynchronous stochastic zeroth-order optimization from $O(\frac{1}{\sqrt{T}})$ to $O(\frac{1}{T})$.

Before providing the theoretical analysis, we give the definitions of Lipschitz constants on the original function and gradient, coordinated smooth function, mixture gradient of the coordinated smooth functions, Lipschitz constant on the mixture gradient, and the explanation of x_t^s used in the analysis as follows, which are critical to the analysis of AsySZO+.

1. **Lipschitz smooth assumptions:** For the smooth functions f_i, we have the Lipschitz constant L for ∇f_i as following.

   Assumption 1 *L is the Lipschitz constant for ∇f_i ($\forall i \in \{1, \cdots, l\}$) in (1). Thus, $\forall x$ and $\forall y$, L-Lipschitz smooth can be presented as*

   $$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\| \tag{18}$$

   Equivalently, L-Lipschitz smooth can also be written as the formulation (19).

   $$f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \tag{19}$$

L_0 is the Lipschitz constant for f_i ($\forall i \in \{1, \cdots, l\}$) in (1). $\forall x$ and $\forall y$, we have that

$$|f_i(x) - \nabla f_i(y)| \leq L_0 \|x - y\| \tag{20}$$

2. **Coordinated smooth function:** Given a function $f(x)$ and a predefined approximation parameter vector $[\mu_1, \mu_2, \cdots, \mu_N]$, we define a coordinated smooth function $f^j(x)$ w.r.t the j-th dimension which was used in (Lian et al., 2016).

$$f^j(x) = \mathbb{E}_{v \sim U_{[-\mu_j, \mu_j]}}(p(x + ve_j)) = \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} f(x + ve_j) dv \tag{21}$$

where $v \sim U_{[-\mu_j, \mu_j]}$ means that v follows the uniform distribution over the interval $[-\mu_j, \mu_j]$. It should be noted that, we have the following equation between $G_j(x, f)$ and $\nabla_j f^j(x)$.

$$\begin{aligned} \nabla_j f^j(x) &= \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \nabla_j f(x + ve_j) dv \\ &= \frac{1}{2\mu_j} \left(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j) \right) e_j = N G_j(x, f) \end{aligned} \tag{22}$$

In addition, we have

$$\mathbb{E}_j \|\nabla_j f^j(x) - \nabla_j f(x)\| \leq \frac{L^2 \sum_{j=1}^N \mu_j^2}{4N} \stackrel{\text{def}}{=} \frac{\omega}{4} \tag{23}$$

which is proved in (26) of (Lian et al., 2016).

3. **Mixture gradient of the coordinated smooth functions:** Based on the coordinated smooth function $f^j(x)$, we define a mixture gradient on the coordinated smooth functions as $\sum_{j=1}^N \nabla_j f^j(x)$.

4. **Lipschitz constant on the mixture gradient:** We assume that there exists a Lipschitz constant ($\widetilde{L}$) on the mixture gradient as follows.

Lemma 1 *There exists a Lipschitz constant $\widetilde{L}$ for the mixture gradient $\sum_{j=1}^N \nabla_j f_i^j(x)$ ($\forall i \in \{1, \ldots, l\}$), such that, $\forall x$ and $\forall y$, we have*

$$\left\| \sum_{j=1}^N \nabla_j f_i^j(x) - \sum_{j=1}^N \nabla_j f_i^j(y) \right\| \leq \widetilde{L} \|x - y\| \tag{24}$$

where $\widetilde{L} \leq \min\{NL, \sum_{j=1}^N \frac{L_0}{\mu_j}\}$.

Proof We first prove that $\widetilde{L} \leq NL$.

$$\left\| \sum_{j=1}^N \nabla_j f^j(x) - \sum_{j=1}^N \nabla_j f^j(y) \right\| \tag{25}$$

$$
\begin{aligned}
&= \left\| \sum_{j=1}^{N} \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \nabla_j f(x + ve_j) dv - \sum_{j=1}^{N} \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \nabla_j f(y + ve_j) dv \right\| \\
&= \left\| \sum_{j=1}^{N} \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \left(\nabla_j f(x + ve_j) - \nabla_j f(y + ve_j) \right) dv \right\| \\
&\leq \sum_{j=1}^{N} \frac{1}{2\mu_j} \left\| \int_{-\mu_j}^{\mu_j} \left(\nabla_j f(x + ve_j) - \nabla_j f(y + ve_j) \right) dv \right\| \\
&\leq \sum_{j=1}^{N} \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \left\| \nabla_j f(x + ve_j) - \nabla_j f(y + ve_j) \right\| dv \\
&\leq \sum_{j=1}^{N} \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \left\| \nabla f(x + ve_j) - \nabla f(y + ve_j) \right\| dv \\
&\leq \sum_{j=1}^{N} \frac{L}{2\mu_j} \int_{-\mu_j}^{\mu_j} \left\| x - y \right\| dv \\
&= NL \left\| x - y \right\|
\end{aligned}
$$

Then, we prove that $\widetilde{L} \leq \sum_{j=1}^{N} \frac{L_0}{\mu_j}$.

$$
\left\| \sum_{j=1}^{N} \nabla_j f^j(x) - \sum_{j=1}^{N} \nabla_j f^j(y) \right\| \tag{26}
$$

$$
\begin{aligned}
&= \left\| \sum_{j=1}^{N} \frac{1}{2\mu_j} \left(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j) \right) e_j - \sum_{j=1}^{N} \frac{1}{2\mu_j} \left(f_i(y + \mu_j e_j) - f_i(y - \mu_j e_j) \right) e_j \right\| \\
&= \left\| \sum_{j=1}^{N} \frac{1}{2\mu_j} \left(\left(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j) \right) - \left(f_i(y + \mu_j e_j) - f_i(y - \mu_j e_j) \right) \right) e_j \right\| \\
&\leq \sum_{j=1}^{N} \frac{1}{2\mu_j} \left\| \left(\left(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j) \right) - \left(f_i(y + \mu_j e_j) - f_i(y - \mu_j e_j) \right) \right) e_j \right\| \\
&= \sum_{j=1}^{N} \frac{1}{2\mu_j} \left\| \left(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j) \right) - \left(f_i(y + \mu_j e_j) - f_i(y - \mu_j e_j) \right) \right\| \\
&\leq \sum_{j=1}^{N} \frac{1}{2\mu_j} \left(\left\| f_i(x + \mu_j e_j) - f_i(y + \mu_j e_j) \right\| + \left\| f_i(x - \mu_j e_j) - f_i(y - \mu_j e_j) \right\| \right) \\
&\leq \sum_{j=1}^{N} \frac{L_0}{2\mu_j} \left(\left\| x - y \right\| + \left\| x - y \right\| \right) \\
&= \sum_{j=1}^{N} \frac{L_0}{\mu_j} \left\| x - y \right\|
\end{aligned}
$$

This completes the proof. ■

Because $f^j(x)$ is a smooth function of $f(x)$, it is reasonable to have a Lipschitz constant on the mixture gradient. Specifically, if $[\mu_1, \mu_2, \cdots, \mu_N] = \mathbf{0}$, it is easy to verify that $\widetilde{L} = L$. If $\mu_j = \infty$ for all $j = 1, \cdots, N$, it is easy to verify that $\widetilde{L} = 0$. Note that, it is possible that $\widetilde{L} > L$.

Correspondingly, we assume there exists a relationship constant $\widehat{L}$ between the original gradient and the mixture gradient, as follows. Note that, it is also possible that $\widehat{L} > 1$.

Assumption 2 *For a smooth function f, we have the relationship constant $\widehat{L}$ between the original gradient and the mixture gradient as*

$$\left\| \sum_{j=1}^{N} \nabla_j f^j(x) \right\| \leq \widehat{L} \|\nabla f(x)\| \tag{27}$$

5. x_t^s**:** As mentioned previously, AsySZO+ does not use any locks in the reading and writing. Thus, in the line 10 of Algorithm 1, x_t^s (left side of '$\leftarrow$') updated in the shared memory may be inconsistent with the ideal one (right side of '$\leftarrow$') computed by the proximal operator. In the analysis, we use x_t^s to denote the ideal one computed by the proximal operator. Same as mentioned in (Mania et al., 2015), there might not be an actual time the ideal ones exist in the shared memory, except the first and last iterates for each outer loop. It is noted that, x_0^s and x_m^s are exactly what is stored in shared memory. Thus, we only consider the ideal x_t^s in the analysis.

Then, we give the upper bounds of $\mathbb{E}\|G(x; f_i) - G(y; f_i)\|^2$ and $\sum_{t=0}^{m-1} \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2$ in Lemma 1.5 and 2 respectively. Based on Lemma 1.5 and 2, we give an upper bound of $\sum_{t=0}^{m-1} \mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2$ (Theorem 1.5). Then, we prove the sublinear rate of the convergence (Theorem 2 and Corollary 1).

Lemma 1.5 *For the smooth function f_i and the corresponding approximate full gradient $G(x; f_i)$, under Assumption 1, we have*

$$\mathbb{E}\|G(x; f_i) - G(y; f_i)\|^2 \leq \widetilde{L}^2 \|x - y\|^2 \tag{28}$$

Proof Based on the definition of the approximate gradient $G(x; f_i)$, we have that

$$\mathbb{E}\|G(x; f_i) - G(y; f_i)\|^2 = \mathbb{E}\left\| \frac{1}{N} \sum_{j=1}^{N} (G_j(x; f_i) - G_j(y; f_i)) \right\|^2 \tag{29}$$

$$= \mathbb{E}\left\| \sum_{j=1}^{N} \left(\nabla_j f_i^j(x) - \nabla_j f_i^j(y) \right) \right\|^2 \leq \widetilde{L}^2 \|x - y\|^2$$

where the second equality uses (22), the first inequality uses (24). This completes the proof. ■

Lemma 2 *If $Y - 2N\widetilde{L}^2\gamma^2\tau^2 > 0$, under Assumptions 1 and 2, we have that*

$$\sum_{t=0}^{m-1} \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 \leq \frac{2Y}{Y - 2N\widetilde{L}^2\gamma^2\tau^2} \sum_{t=0}^{m-1} \left(\frac{2N\widetilde{L}^2}{b} \left\|x_t^{s+1} - \widetilde{x}^s\right\|^2 + 2\widehat{L}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 \right) \tag{30}$$

Proof Let $v_t^{s+1} = \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(x_t^{s+1}; f_i) - \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(\widetilde{x}^s; f_i) + G(\widetilde{x}^s; f)$, we have that

$$\begin{aligned} & \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 = \mathbb{E}\left\|\widehat{v}_t^{s+1} - v_t^{s+1} + v_t^{s+1}\right\|^2 \qquad (31) \\ \leq\ & 2\mathbb{E}\left\|\widehat{v}_t^{s+1} - v_t^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ =\ & 2\mathbb{E}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)} \left(G(\widehat{x}_t^{s+1}; f_i) - G(x_t^{s+1}; f_i)\right)\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ \leq\ & \frac{2}{b}\sum_{i\in\mathcal{B}(t)} \mathbb{E}\left\|G(\widehat{x}_t^{s+1}; f_i) - G(x_t^{s+1}; f_i)\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ \leq\ & 2\widetilde{L}^2\mathbb{E}\left\|\widehat{x}_t^{s+1} - x_t^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ =\ & 2\widetilde{L}^2\gamma^2\mathbb{E}\left\|\sum_{t'\in K(t)} B_{t'}^{s+1}\widehat{v}_{J(t')}^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ \leq\ & 2\widetilde{L}^2\gamma^2\tau\mathbb{E}\sum_{t'\in K(t)} \left\|B_{t'}^{s+1}\widehat{v}_{J(t')}^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ \leq\ & 2\widetilde{L}^2\gamma^2\tau\sum_{t'\in K(t)} \mathbb{E}\left\|\widehat{v}_{J(t')}^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \\ =\ & \frac{2N\widetilde{L}^2\gamma^2\tau}{Y}\sum_{t'\in K(t)} \mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \end{aligned}$$

where the first, second and fourth inequalities use the fact that $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$, the third inequality uses (28), the fifth inequality uses the Cauchy-Schwarz inequality and the fact $\left\|B_t^{s+1}\right\| \leq 1$. We consider a fixed stage $s+1$ such that $x_0^{s+1} = x_m^s$. By summing the the inequality (31) over $t = 0, \cdots, m-1$, we obtain

$$\begin{aligned} \sum_{t=0}^{m-1} \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 &\leq \sum_{t=0}^{m-1} \left(\frac{2N\widetilde{L}^2\gamma^2\tau}{Y} \sum_{t'\in K(t)} \mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2 + 2\mathbb{E}\left\|v_t^{s+1}\right\|^2 \right) \\ &\leq \frac{2N\widetilde{L}^2\gamma^2\tau^2}{Y} \sum_{t=0}^{m-1} \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 + 2\sum_{t=0}^{m-1} \mathbb{E}\left\|v_t^{s+1}\right\|^2 \end{aligned} \tag{32}$$

where the second inequality uses the Assumption 1. If $Y - 2N\widetilde{L}^2\gamma^2\tau^2 > 0$, we have that

$$\sum_{t=0}^{m-1} \mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 \leq \frac{2Y}{Y - 2N\widetilde{L}^2\gamma^2\tau^2} \sum_{t=0}^{m-1} \mathbb{E}\left\|v_t^{s+1}\right\|^2 \tag{33}$$

We next bound $\mathbb{E}\left\|v_t^{s+1}\right\|^2$ by

$$\begin{aligned}
& \mathbb{E}\left\|v_t^{s+1}\right\|^2 \qquad (34) \\
= \ & \mathbb{E}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(x_t^{s+1};f_i) - \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(\widetilde{x}^s;f_i) + G(\widetilde{x}^s;f)\right\|^2 \\
= \ & \mathbb{E}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(x_t^{s+1};f_i) - \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(\widetilde{x}^s;f_i) + G(x^s;f) - G(x_t^{s+1};f) + G(\widetilde{x}_t^{s+1};f)\right\|^2 \\
\leq \ & 2\mathbb{E}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(x_t^{s+1};f_i) - \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(\widetilde{x}^s;f_i) - \big(G(x_t^{s+1};f) - G(\widetilde{x}^s;f)\big)\right\|^2 + 2\mathbb{E}\left\|G(x_t^{s+1};f)\right\|^2 \\
= \ & \frac{2}{b^2}\mathbb{E}\left\|\sum_{i\in\mathcal{B}(t)} \Big(G(x_t^{s+1};f_i) - G(\widetilde{x}^s;f_i) - \big(G(x_t^{s+1};f) - G(\widetilde{x}^s;f)\big)\Big)\right\|^2 + 2\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N} G_j(x_t^{s+1};f)\right\|^2 \\
\leq \ & \frac{2}{b}\mathbb{E}\left\|G(x_t^{s+1};f_i) - G(\widetilde{x}^s;f_i) - G(x_t^{s+1};f) + G(\widetilde{x}^s;f)\right\|^2 + 2\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N} G_j(x_t^{s+1};f)\right\|^2 \\
\leq \ & \frac{2}{b}\mathbb{E}\left\|G(x_t^{s+1};f_i) - G(\widetilde{x}^s;f_i)\right\|^2 + 2\mathbb{E}\left\|\sum_{j=1}^{N} \nabla_j f^j(x_t^{s+1})\right\|^2 \\
\leq \ & \frac{2\widetilde{L}^2}{b}\left\|x_t^{s+1} - \widetilde{x}^s\right\|^2 + 2\widehat{L}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2
\end{aligned}$$

where the first inequality uses $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$, The second inequality uses Lemma 7 in (Reddi et al., 2016), the third inequality uses $\mathbb{E}\|x-\mathbb{E}x\|^2 \leq \mathbb{E}\|x\|^2$, the fourth inequality uses (28) and (27). This completes the proof. ■

Theorem 1.5 *Setting $c_m = 0$, $\beta_t > 0$. Let*

$$c_t = c_{t+1}(1+\gamma\beta_t) + \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3 NL^2\tau^2}{Y}\right)\frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)} \qquad (35)$$

$$\Gamma_t = \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3 NL^2\tau^2}{Y}\right)\frac{4Y\widehat{L}}{Y-2N\widetilde{L}^2\gamma^2\tau^2} \qquad (36)$$

Let η_t, β_t and c_{t+1} be chosen such that $\Gamma_t > 0$ and $\beta_t \geq 2c_{t+1}$. Under Assumptions 1, 1, 1 and 2, $\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2$ in AsySZO+ satisfy the bound

$$\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 \leq \frac{\mathbb{E}(f(x^s)) - \mathbb{E}(f(x^{s+1})) + \frac{\gamma N\omega m}{4}}{\min_{t\in\{0,\cdots,m-1\}}\Gamma_t} \qquad (37)$$

Proof We first bound $\mathbb{E}\left\|x_{t+1}^{s+1}-\widetilde{x}^s\right\|^2$.

$$
\begin{aligned}
& \mathbb{E}\left\|x_{t+1}^{s+1}-\widetilde{x}^s\right\|^2=\mathbb{E}\left\|x_{t+1}^{s+1}-x_t^{s+1}+x_t^{s+1}-\widetilde{x}^s\right\|^2 \qquad (38)\\
= \; & \mathbb{E}\left(\left\|x_{t+1}^{s+1}-x_t^{s+1}\right\|^2+\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+2\left\langle x_{t+1}^{s+1}-x_t^{s+1}, x_t^{s+1}-\widetilde{x}^s\right\rangle\right) \\
= \; & \mathbb{E}\left(\gamma^2\left\|\widehat{v}_{J(t)}^{s+1}\right\|^2+\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2-2\gamma\left\langle\widehat{v}_{J(t)}^{s+1}, x_t^{s+1}-\widetilde{x}^s\right\rangle\right) \\
= \; & \frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2-2\gamma\mathbb{E}\left\langle\frac{1}{b}\sum_{i\in\mathcal{B}(t)}G(\widehat{x}_t^{s+1};f_i), x_t^{s+1}-\widetilde{x}^s\right\rangle \\
\leq \; & \frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+2\gamma\mathbb{E}\left(\frac{1}{2\beta_t}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)}G(\widehat{x}_t^{s+1};f_i)\right\|^2+\frac{\beta_t}{2}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2\right) \\
= \; & \frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+2\gamma\mathbb{E}\left(\frac{1}{2\beta_t}\left\|\frac{1}{b}\sum_{i\in\mathcal{B}(t)}\sum_{j=1}^N\nabla_j f^j(x_t^{s+1})\right\|^2\right) \\
\leq \; & \frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma}{b\beta_t}\mathbb{E}\left(\sum_{i\in\mathcal{B}(t)}\left\|\sum_{j=1}^N\nabla_j f^j(x_t^{s+1})\right\|^2\right) \\
= \; & \frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N}{\beta_t}\mathbb{E}\left\|\nabla_j f^j(x_t^{s+1})\right\|^2
\end{aligned}
$$

where the first inequality uses the Young's inequality, the second inequality uses the fact that $\|\sum_{i=1}^n a_i\|^2\leq n\sum_{i=1}^n\|a_i\|^2$. We next bound $\mathbb{E}\left\|\nabla_j f(x_t^{s+1})-\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2$.

$$
\begin{aligned}
& \mathbb{E}\left\|\nabla_j f(x_t^{s+1})-\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2 \qquad (39)\\
= \; & \mathbb{E}\left\|\nabla_j f(x_t^{s+1})-\nabla_j f(\widehat{x}_t^{s+1})+\nabla_j f(\widehat{x}_t^{s+1})-\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2 \\
\leq \; & 2\mathbb{E}\left\|\nabla_j f(x_t^{s+1})-\nabla_j f(\widehat{x}_t^{s+1})\right\|^2+2\mathbb{E}\left\|\nabla_j f(\widehat{x}_t^{s+1})-\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2 \\
\leq \; & \frac{2}{N}\mathbb{E}\left\|\nabla f(x_t^{s+1})-\nabla f(\widehat{x}_t^{s+1})\right\|^2+\frac{\omega}{2} \\
\leq \; & \frac{2L^2}{N}\left\|x_t^{s+1}-\widehat{x}_t^{s+1}\right\|^2+\frac{\omega}{2} \\
= \; & \frac{2L^2\gamma^2}{N}\left\|\sum_{t'\in K(t)}B_{t'}^{s+1}\widehat{v}_{J(t')}^{s+1}\right\|^2+\frac{\omega}{2} \\
\leq \; & \frac{2L^2\gamma^2\tau}{N}\mathbb{E}\sum_{t'\in K(t)}\left\|B_{t'}^{s+1}\widehat{v}_{J(t')}^{s+1}\right\|^2+\frac{\omega}{2} \\
\leq \; & \frac{2L^2\gamma^2\tau}{N}\mathbb{E}\sum_{t'\in K(t)}\left\|\widehat{v}_{J(t')}^{s+1}\right\|^2+\frac{\omega}{2} \\
= \; & \frac{2L^2\gamma^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2+\frac{\omega}{2}
\end{aligned}
$$

where the first and fourth inequalities use $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$, the second inequality uses (23), the third inequality uses (20), the fifth inequality uses the Cauchy-Schwarz inequality and the fact $\left\|B_t^{s+1}\right\| \leq 1$. We bound $\mathbb{E}\left(f(x_{t+1}^{s+1})\right)$ as follows.

$$
\begin{aligned}
& \mathbb{E}\left(f(x_{t+1}^{s+1})\right) && (40)\\
\leq\ & \mathbb{E}\left(f(x_t^{s+1}) + \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1}\rangle + \frac{L}{2}\left\|x_{t+1}^{s+1} - x_t^{s+1}\right\|^2\right)\\
=\ & \mathbb{E}\left(f(x_t^{s+1}) - \gamma\left\langle \nabla f(x_t^{s+1}), \widehat{v}_{J(t)}^{s+1}\right\rangle + \frac{L\gamma^2}{2}\left\|\widehat{v}_{J(t)}^{s+1}\right\|^2\right)\\
=\ & \mathbb{E}f(x_t^{s+1}) - \gamma\mathbb{E}\left\langle \nabla f(x_t^{s+1}), \frac{1}{b}\sum_{i\in\mathcal{B}(t)} G(\widehat{x}_t^{s+1}; f_i)\right\rangle + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2\\
=\ & \mathbb{E}f(x_t^{s+1}) - \gamma\mathbb{E}\left\langle \nabla f(x_t^{s+1}), \frac{1}{N}\sum_{j=1}^{N} G_j(\widehat{x}_t^{s+1}; f)\right\rangle + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2\\
=\ & \mathbb{E}f(x_t^{s+1}) - \gamma\mathbb{E}\left\langle \nabla f(x_t^{s+1}), \sum_{j=1}^{N} \nabla_j f^j(\widehat{x}_t^{s+1})\right\rangle + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2\\
=\ & \mathbb{E}f(x_t^{s+1}) + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2\\
& -\frac{\gamma}{2}\left(\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 + \mathbb{E}\left\|\sum_{j=1}^{N}\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2 - \mathbb{E}\left\|\nabla f(x_t^{s+1}) - \sum_{j=1}^{N}\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\right)\\
=\ & \mathbb{E}f(x_t^{s+1}) + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 - \frac{\gamma N}{2}\mathbb{E}\left\|\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\\
& +\frac{\gamma N}{2}\mathbb{E}\left\|\nabla_j f(x_t^{s+1}) - \nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\\
\leq\ & \mathbb{E}f(x_t^{s+1}) + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 - \frac{\gamma N}{2}\mathbb{E}\left\|\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\\
& +\frac{\gamma^3 N L^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2 + \frac{\gamma N\omega}{4}
\end{aligned}
$$

where the first inequality uses (19), the second inequality uses (39). Next, we define Lyapunov function $R_t^{s+1} = \mathbb{E}\left(f(x_t^{s+1}) + c_t\left\|x_t^{s+1} - \widetilde{x}^s\right\|^2\right)$, and give the upper bound of R_{t+1}^{s+1} as follows.

$$
\begin{aligned}
& R_{t+1}^{s+1} && (41)\\
=\ & \mathbb{E}\left(f(x_{t+1}^{s+1}) + c_{t+1}\left\|x_{t+1}^{s+1} - \widetilde{x}^s\right\|^2\right)\\
\leq\ & \mathbb{E}f(x_t^{s+1}) + \frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 - \frac{\gamma N}{2}\mathbb{E}\left\|\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\\
& +\frac{\gamma^3 N L^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2 + \frac{\gamma N\omega}{4}
\end{aligned}
$$

$$
\begin{aligned}
&+c_{t+1}\left(\frac{N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N}{\beta_t}\mathbb{E}\left\|\nabla_j f^j(x_t^{s+1})\right\|^2\right)\\
= & \ \mathbb{E}f(x_t^{s+1})+\frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2-\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2-\left(\frac{\gamma N}{2}-\frac{c_{t+1}\gamma N}{\beta_t}\right)\mathbb{E}\left\|\nabla_j f^j(\widehat{x}_t^{s+1})\right\|^2\\
&+\frac{\gamma^3 NL^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2+\frac{c_{t+1}N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+c_{t+1}(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N\omega}{4}\\
\leq & \ \mathbb{E}f(x_t^{s+1})+\frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2-\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2+\frac{\gamma^3 NL^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2\\
&+\frac{c_{t+1}N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+c_{t+1}(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N\omega}{4}
\end{aligned}
$$

where the first inequality uses (38) and (40), and the second inequality uses the constraint $\beta_t \geq 2c_{t+1}$. We consider a fixed stage $s+1$ such that $x_0^{s+1} = x_m^s$. By summing the the inequality (41) over $t = 0, \cdots, m-1$, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{m-1} R_{t+1}^{s+1} \qquad (42)\\
\leq & \sum_{t=0}^{m-1}\Bigg(\mathbb{E}f(x_t^{s+1})+\frac{LY\gamma^2}{2N}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2-\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2+\frac{\gamma^3 NL^2\tau}{Y}\sum_{t'\in K(t)}\mathbb{E}\left\|\widehat{v}_{t'}^{s+1}\right\|^2\\
&+\frac{c_{t+1}N\gamma^2}{Y}\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2+c_{t+1}(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N\omega}{4}\Bigg)\\
= & \sum_{t=0}^{m-1}\Bigg(\mathbb{E}f(x_t^{s+1})-\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2+c_{t+1}(1+\gamma\beta_t)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2+\frac{\gamma N\omega}{4}\\
&+\left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\mathbb{E}\left\|\widehat{v}_t^{s+1}\right\|^2\Bigg)\\
\leq & \sum_{t=0}^{m-1}\Bigg(\mathbb{E}f(x_t^{s+1})+\frac{\gamma N\omega}{4}\\
&-\left(\frac{\gamma}{2}-\left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4Y\widehat{L}}{Y-2N\widetilde{L}^2\gamma^2\tau^2}\right)\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2+\\
&\left(c_{t+1}(1+\gamma\beta_t)+\left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}\right)\mathbb{E}\left\|x_t^{s+1}-\widetilde{x}^s\right\|^2\Bigg)\\
= & \sum_{t=0}^{m-1}\left(R_t^{s+1}-\Gamma_t\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2+\frac{\gamma N\omega}{4}\right)
\end{aligned}
$$

where the second inequality uses (30). Because $c_m = 0$, we have that $R_m^{s+1} = \mathbb{E}(f(x_m^{s+1})) = \mathbb{E}(f(x^{s+1}))$. In addition, we have that $R_0^{s+1} = \mathbb{E}(f(x_0^{s+1})) = \mathbb{E}(f(x^s))$. Based on (42), we have that

$$\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 \leq \frac{\sum_{t=0}^{m-1}(R_t^{s+1}-R_{t+1}^{s+1})+\frac{\gamma N\omega m}{4}}{\min_{t\in\{0,\cdots,m-1\}}\Gamma_t} = \frac{(R_0^{s+1}-R_m^{s+1})+\frac{\gamma N\omega m}{4}}{\min_{t\in\{0,\cdots,m-1\}}\Gamma_t} = \frac{\mathbb{E}(f(x^s))-\mathbb{E}(f(x^{s+1}))+\frac{\gamma N\omega m}{4}}{\min_{t\in\{0,\cdots,m-1\}}\Gamma_t} \tag{43}$$

This completes the proof. ■

Theorem 2 *Let* $c_m=0$, $\gamma=\frac{u_0 b}{\widetilde{L}l^\alpha}$, $\beta_t=\frac{\widetilde{L}N^2}{Y}$, $0<\alpha<1$, $0<u_0<1$, $c_t=c_{t+1}(1+\gamma\beta_t)+\left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}$ *for* $t=0,\cdots,m-1$, *and* $\sigma=u_0\left(\frac{1}{2}-\left(\frac{13LYu_0b}{5N\widetilde{L}}+\frac{8NL^2\tau^2u_0^2b^2}{5Y\widetilde{L}}\right)4\widehat{L}\right)$. *Under Assumptions 1, 1, 1 and 2, if*

$$\tau<\min\left\{\frac{\frac{5\widetilde{L}N^2}{2Y}-\frac{2LY^2}{N^2}}{8L^2u_0b},\frac{\left(\frac{1}{8\widehat{L}}-\frac{13LYu_0b}{5N\widetilde{L}}\right)5Y\widetilde{L}}{8NL^2u_0^2b^2}\right\} \tag{44}$$

, $\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2$ *in AsySZO+ satisfy the bound*

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2\leq\frac{\widetilde{L}l^\alpha\left(f(x^0)-\mathbb{E}(f(x^S))\right)}{\sigma bT}+\frac{Nu_0\omega}{4\sigma} \tag{45}$$

Proof Based on the specified values of γ and β_t, we have that

$$\theta = \gamma\beta_t+\frac{4N^2\gamma^2\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}=\frac{u_0b}{\frac{Yl^\alpha}{N^2}}+\frac{4u_0^2b}{\frac{Yl^{2\alpha}}{N^2}-\frac{2\tau^2u_0^2b^2}{N}} = \frac{u_0bN^2}{Yl^\alpha}+\frac{4u_0^2bN^2}{Yl^{2\alpha}-2N\tau^2u_0^2b^2} \leq \frac{5u_0bN^2}{Yl^\alpha} \tag{46}$$

where the inequality uses the constraint $Yl^\alpha\leq Yl^{2\alpha}-2N\tau^2u_0^2b^2$ by appropriately choosing α and u_0. We set $m=\lfloor\frac{Yl^\alpha}{5u_0bN^2}\rfloor$, from the recurrence definition of c_t, we have that

$$c_0 = \frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}\left(\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{(1+\theta)^m-1}{\theta} = \frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}\frac{\frac{LYu_0^2b^2}{2N\widetilde{L}^2l^{2\alpha}}+\frac{NL^2\tau^2u_0^3b^3}{Y\widetilde{L}^3l^{3\alpha}}}{\frac{u_0bN^2}{Yl^\alpha}+\frac{4u_0^2bN^2}{Yl^{2\alpha}-2N\tau^2u_0^2b^2}}((1+\theta)^m-1) \tag{47}$$

$$
\begin{aligned}
&\leq \frac{4YN\widetilde{L}^2 l^{2\alpha}}{b(Yl^{2\alpha}-2N\tau^2u_0^2b^2)} \frac{\frac{LYu_0^2b^2}{2N}+\frac{NL^2\tau^2u_0^3b^3}{Y}\left(Yl^{2\alpha}-2N\tau^2u_0^2b^2\right)}{5u_0^2bN^2\widetilde{L}^2l^{2\alpha}}\left((1+\theta)^m-1\right) \\
&= \frac{\frac{2LY^2}{N}+4NL^2\tau^2u_0b}{5N}\left((1+\theta)^m-1\right) \\
&\leq \frac{\frac{2LY^2}{N}+4NL^2\tau^2u_0b}{5N}(e-1)
\end{aligned}
$$

where the first inequality uses $\widetilde{L}^3l^{3\alpha} \geq \widetilde{L}^2l^{2\alpha}$, the second inequality uses the fact $(1+\frac{1}{a})^a$ is increasing for $a>0$, and $\lim_{a\to\infty}(1+\frac{1}{a})^a = e$, which is also used in (Reddi et al., 2015). Because $\beta_t \geq 2c_{t+1}$, we have that $\tau^2 \leq \frac{\frac{5\widetilde{L}N^2}{2Y}-\frac{2LY^2}{N^2}}{8L^2u_0b}$ from (47). Let $\widetilde{\Gamma}$ denote the following quantity:

$$
\widetilde{\Gamma} = \min_{t\in\{0,\cdots,m-1\}} \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4Y\widehat{L}}{Y-2N\widetilde{L}^2\gamma^2\tau^2} \tag{48}
$$

Now we give a lower bound of $\widetilde{\Gamma}$ as

$$
\begin{aligned}
\widetilde{\Gamma} &= \min_{t\in\{0,\cdots,m-1\}} \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4Y\widehat{L}}{Y-2N\widetilde{L}^2\gamma^2\tau^2} \qquad (49) \\
&\geq \frac{\gamma}{2} - \left(\frac{c_0N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4Y\widehat{L}}{Y-2N\widetilde{L}^2\gamma^2\tau^2} \\
&\geq \frac{\gamma}{2} - \left(\frac{c_0N\gamma^2}{Y}+\frac{LY\gamma^2}{2N}+\frac{\gamma^3NL^2\tau^2}{Y}\right)4\widehat{L}l^{\alpha} \\
&\overset{(47)}{\geq} \frac{\gamma}{2} - \left(\frac{\frac{\frac{2LY^2}{N}+4NL^2\tau^2u_0b}{5N}(e-1)N\gamma}{Y}+\frac{LY\gamma}{2N}+\frac{\gamma^2NL^2\tau^2}{Y}\right)4\gamma\widehat{L}l^{\alpha} \\
&= \frac{\gamma}{2} - \left(\frac{2LY(e-1)u_0b}{5N\widetilde{L}}+\frac{4NL^2\tau^2u_0^2b^2(e-1)}{5Y\widetilde{L}}+\frac{LYu_0b}{2N\widetilde{L}}+\frac{u_0^2b^2NL^2\tau^2}{Y\widetilde{L}^2l^{\alpha}}\right)4\gamma\widehat{L} \\
&= \gamma\left(\frac{1}{2} - \left(\frac{(4(e-1)+5)LYu_0b}{5N\widetilde{L}}+\frac{NL^2\tau^2u_0^2b^2}{Y\widetilde{L}}\left(\frac{4(e-1)}{5}+\frac{1}{\widetilde{L}l^{\alpha}}\right)\right)4\widehat{L}\right) \\
&\geq \underbrace{\gamma\left(\frac{1}{2} - \left(\frac{13LYu_0b}{5N\widetilde{L}}+\frac{8NL^2\tau^2u_0^2b^2}{5Y\widetilde{L}}\right)4\widehat{L}\right)}_{\varrho} \\
&= \frac{\sigma b}{\widetilde{L}l^{\alpha}}
\end{aligned}
$$

where the first inequality holds because c_t decrease with t, the second inequality uses the constraint $Yl^{\alpha} \leq Yl^{2\alpha} - 2N\tau^2u_0^2b^2$, the fourth inequality uses the constraint $\widetilde{L}l^{\alpha} \geq \frac{5}{8-4(e-1)}$. For the last inequality, we can appropriately choose a value of u_0, such that $\varrho > 0$, and σ is a small value independent to l, specifically $\sigma = u_0\left(\frac{1}{2} - \left(\frac{13LYu_0b}{5N\widetilde{L}}+\frac{8NL^2\tau^2u_0^2b^2}{5Y\widetilde{L}}\right)4\widehat{L}\right)$.

Because $\varrho > 0$, we have that $\tau^2 < \frac{\left(\frac{1}{8\widehat{L}} - \frac{13LYu_0b}{5N\widetilde{L}}\right)5Y\widetilde{L}}{8NL^2u_0^2b^2}$. Based on (49), we have that

$$\begin{aligned} \frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 &\leq \frac{1}{T}\sum_{s=0}^{S-1}\frac{\mathbb{E}(f(x^s)) - \mathbb{E}(f(x^{s+1})) + \frac{\gamma N\omega m}{4}}{\widetilde{\Gamma}} \\ &= \frac{f(x^0)) - \mathbb{E}(f(x^S)) + \frac{\gamma N\omega T}{4}}{T\widetilde{\Gamma}} \\ &\leq \frac{\widetilde{L}l^\alpha\left(f(x^0)) - \mathbb{E}(f(x^S))\right)}{\sigma bT} + \frac{Nu_0\omega}{4\sigma} \end{aligned} \quad (50)$$

This completes the proof. ∎

Corollary 1 *Let* $c_m = 0$, $\gamma = \frac{u_0b}{\widetilde{L}l^\alpha}$, $\beta_t = \frac{\widetilde{L}N^2}{Y}$, $0 < \alpha < 1$, $0 < u_0 < 1$, *and* $c_t = c_{t+1}(1+\gamma\beta_t) + \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y}\right)\frac{4YN\widetilde{L}^2}{b(Y-2N\widetilde{L}^2\gamma^2\tau^2)}$ *for* $t = 0,\cdots,m-1$, $\sigma = u_0\left(\frac{1}{2} - \left(\frac{13LYu_0b}{5N\widetilde{L}} + \frac{8NL^2\tau^2u_0^2b^2}{5Y\widetilde{L}}\right)4\widehat{L}\right)$. *Under Assumptions 1, 1, 1 and 2, if* $\omega = 0$,

$$\tau < \min\left\{\frac{\frac{5\widetilde{L}N^2}{2Y} - \frac{2LY^2}{N^2}}{8L^2u_0b}, \frac{\left(\frac{1}{8\widehat{L}} - \frac{13LYu_0b}{5N\widetilde{L}}\right)5Y\widetilde{L}}{8NL^2u_0^2b^2}\right\} \quad (51)$$

, $\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2$ *in AsySZO+ satisfy the bound*

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\left\|\nabla f(x_t^{s+1})\right\|^2 \leq \frac{\widetilde{L}l^\alpha\left(f(x^0)) - \mathbb{E}(f(x^S))\right)}{\sigma bT} \quad (52)$$

References

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *arXiv preprint arXiv:1606.00498*, 2016.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems*, pages 2647–2655, 2015.

Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alex Smola. Fast stochastic methods for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1605.06900*, 2016.