Appendix

In appendix, we provide the proof of our theorems in the paper. In all theorems, we use unbiased stochastic gradients to update the optimization.

1 Proof of theorem 4

Theorem. *Suppose* $f \in \mathcal{F}_n$ *have* σ*-bounded gradient. Let* $\eta_t = \eta_{\Delta_{\text{unbiased}}} = c_{\text{unbiased}}/\sqrt{\Delta+1}$ *for* $0 \le \Delta \le T-1$ *where* $c_{\text{unbiased}} = \sqrt{\frac{f(x_0)-f(x^*)}{(2\lambda^2-2\lambda+1)L\sigma^2}}$ *and let* T *be a multiple of* m*. Further let* $p_m = 1$*, and* $p_i = 0$ *for* $0 \le i < m$*. Then the output* x_a *of Alg. 2 we have*

$$\mathbb{E}[\| \nabla f(x_a)^2 \|] \le \frac{\sqrt{(2\lambda^2-2\lambda+1)}}{(1-\lambda)}\sqrt{\frac{2(f(x^0)-f(x^*))L}{T}}\sigma$$

Proof. As the learning rate decay from 1 to T, we use Definition 2 to bound gradients v_t^{s+1} as following:

$$\begin{aligned}
&\mathbb{E}[\| v_t^{s+1} \|^2] \\
&= \mathbb{E}[\| (1-\lambda)\nabla f_{i_t}(x_t^{s+1}) - \lambda(\nabla f_{i_t}(\tilde{x}^s) - \nabla f(\tilde{x}^s) \|^2]) \\
&\le 2(\mathbb{E}[(\| (1-\lambda)\nabla f_{i_t}(x_t^{s+1}) \|^2 + \| \lambda(\nabla f_{i_t}(\tilde{x}^s) - \nabla f(\tilde{x}^s)) \|^2]) \\
&\le 2((1-\lambda)^2\mathbb{E}[\| \nabla f_{i_t}(x_t^{s+1}) \|^2] + \lambda^2\mathbb{E}[\| \nabla f_{i_t}(\tilde{x}^s) \|^2]) \\
&\le (4\lambda^2-4\lambda+2)\sigma^2,
\end{aligned} \tag{1}$$

where the first inequality we followed Lemma 3 when r=2. The second inequality we followed (a) σ-bounded gradient property of f and (b) the fact that for a random variable ζ followed $\mathbb{E}[\| \zeta - \mathbb{E}[\zeta] \|^2] \le \mathbb{E}[\| \zeta \|^2]$.

Since f is $\mathcal{L}$-smooth, we have

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1}^{s+1})] \le \mathbb{E}[f(x_t^{s+1}) + \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1} \rangle \\
&+ \frac{L}{2} \| x_{t+1}^{s+1} - x_t^{s+1} \|^2].
\end{aligned} \tag{2}$$

Using Alg. 2 to update and since $\mathbb{E}[\nabla f(x_t^{s+1})] = \nabla f(x_t^{s+1})$ (unbiasedness of the stochastic gradients), Ineq. 2 would be updated as:

$$\mathbb{E}[f(x_{t+1}^{s+1})] \le \mathbb{E}[f(x_t^{s+1}) - \eta_\Delta(1-\lambda) \| \nabla f(x_t^{s+1}) \|^2 + \frac{L\eta_\Delta^2}{2} \| v_t^{s+1} \|^2]. \tag{3}$$

Adding the bound of v_t^{s+1} from Ineq. 1 to Ineq. 3, we can obtain that:

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1}^{s+1})] \le \mathbb{E}[f(x_t^{s+1})] - \eta_\Delta(1-\lambda)\mathbb{E}[\| \nabla f(x_t^{s+1}) \|^2] + \frac{L\eta_\Delta^2}{2}\mathbb{E}[\| v_t^{s+1} \|^2]. \\
&\le \mathbb{E}[f(x_t^{s+1})] - \eta_\Delta(1-\lambda)\mathbb{E}[\| \nabla f(x)_t^{s+1} \|^2] + \frac{L\eta_\Delta^2}{2}(4\lambda^2-4\lambda+2)\sigma^2
\end{aligned} \tag{4}$$

Thus the Ineq. 4 can be alternated as

$$\mathbb{E}[\| \nabla f(x)_t^{s+1} \|^2] \le \frac{1}{\eta_\Delta(1-\lambda)}\mathbb{E}[f(x_t^{s+1}) - f(x_{t+1}^{s+1})] + \frac{L\eta_\Delta(2\lambda^2-2\lambda+1)}{(1-\lambda)}\sigma^2, \tag{5}$$

where $t \in \{0, ..., m-1\}$, $s \in \{0, ..., S-1\}$, $\Delta \in \{0, ..., T-1\}$, and $T = mS$.

The minimum upper bound in Ineq. 6 can be achieved when $t = m-1$ and $s = S-1$, and use the constant η we can obtain:

$$\begin{aligned}
&\min_{t,s}\mathbb{E}[\| \nabla f(x_t^{s+1}) \|^2] \le \frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\| f(x_t^{s+1}) \|^2] + \frac{L\eta(2\lambda^2-2\lambda+1)}{(1-\lambda)}\sigma^2 \\
&\le \frac{1}{T}\left(\frac{1}{\eta(1-\lambda)}\mathbb{E}[f(x^0) - f(x^T)]\right) + \frac{L\eta(2\lambda^2-2\lambda+1)}{(1-\lambda)}\sigma^2 \\
&\le \frac{1}{T\eta(1-\lambda)}(f(x^0) - f(x^*)) + \frac{L\eta(2\lambda^2-2\lambda+1)}{(1-\lambda)}\sigma^2
\end{aligned} \tag{6}$$

The first inequality can hold due to the minimum is less than average. The second inequality is achieved from Eq 5, and the third one is followed the fact that $f(x^*) \le f(x^T)$. To calculate learning rate η, we take the derivative of the last inequality in Inequality 6 as

$$\frac{\partial\left(\frac{1}{T\eta(1-\lambda)}(f(x^0)-f(x^*)) + \frac{L\eta(2\lambda^2-2\lambda+1)}{(1-\lambda)}\sigma^2\right)}{\partial\eta} = 0 \tag{7}$$

Thus, $\eta_{\Delta_{\text{unbiased}}} = \eta = c/\sqrt{\Delta+1}$, where $c_{\text{unbiased}} = \sqrt{\frac{f(x^0)-f(x^*)}{(2\lambda^2-2\lambda+1)L\sigma^2}}$. Bring the result of $\eta_{\Delta_{\text{unbiased}}} = \eta = c_{\text{unbiased}}/\sqrt{\Delta+1}$ to Eq. 6, we can achieve the upper bound of expectation as

$$\min_{t,s}\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{1}{T(1-\lambda)}(\frac{\sqrt{T}(f(x^0)-f(x^*))}{c_{unbiased}}) + \frac{Lc_{unbiased}\sigma}{(1-\lambda)})$$
$$\leq \frac{1}{\sqrt{T}(1-\lambda)}(\frac{1}{c_{unbiased}}(f(x^0)-f(x^*)) + Lc_{unbiased}\sigma). \quad (8)$$

□

For the case that the learning rate depends on the data size n, we provide one useful lemma in Lemma 1 firstly that can be used for proofing our Theorems.

Lemma 1. *For* $c_{t_{unbiased}}, c_{t+1}, \beta_t > 0$, *we have*

$$c_{t_{unbiased}} = c_{t+1}(1+\eta_t\beta_t(1-\lambda)+2(1-\lambda)^2\eta_t^2L^2)+L^3\eta_t^2.$$

Let η_t, β_t *and* c_{t+1} *is given so that the* $\Omega_{t_{unbiased}} > 0$ *can be showed as*

$$\Omega_{t_{unbiased}} = \eta_t - \frac{c_{t+1}\eta_t(1-\lambda)}{\beta_t} - (1-\lambda)^2L\eta_t^2 - 2(1-\lambda)^4c_{t+1}\eta_t^2$$

Thus, the iterates in Alg. 2 satisfy the bound:

$$\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{R_t^{s+1}-R_{t+1}^{s+1}}{\Omega_{t_{unbiased}}}$$

where $R_{t_{unbiased}}^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_{t_{unbiased}}\|(1-\lambda)x_t^{s+1}-\lambda\tilde{x}^s\|^2]$ *for* $0 \leq s \leq S-1$.

Proof. To further bound the result in Ineq. 26 since f is $\mathcal{L}$-smooth, we require to bound the intermediate iterates v_t^{s+1}, which is showed following inequalities:

$$\begin{aligned}
&\mathbb{E}[\|v_t^{s+1}\|^2] \\
&= \mathbb{E}[\|(1-\lambda)(\nabla f_{i_t}(x_t^{s+1}) - \lambda(\nabla f_{i_t}(\tilde{x}^s) - \nabla f(\tilde{x}^s))\|^2]) \\
&= \mathbb{E}[\|\zeta_t^{s+1} + \lambda\nabla f(\tilde{x}^s) - (1-\lambda)\nabla f(x_t^{s+1}) + (1-\lambda)\nabla f(x_t^{s+1})\|^2] \\
&\leq 2\mathbb{E}[\|(1-\lambda)\nabla f(x_t^{s+1})\|^2] + 2\mathbb{E}[\|\zeta_t^{s+1} - \mathbb{E}[\zeta_t^{s+1}]\|^2] \\
&\leq 2(1-\lambda)^2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + 2\mathbb{E}[\|(1-\lambda)\nabla f_{i_t}(x_t^{s+1}) - \lambda\nabla f_{i_t}(\tilde{x}^s)\|^2] \\
&\leq 2(1-\lambda)^2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + 2L^2\mathbb{E}[\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2],
\end{aligned} \quad (9)$$

where $0 \leq \lambda \leq 1$. In the first inequality, the variable ζ is showed as

$$\zeta_t^{s+1} = \frac{1}{|I_t|}\sum_{i_t\in I_t}((1-\lambda)\nabla f_{i_t}(x_t^{s+1}) - \lambda\nabla f_{i_t}(\tilde{x}^s)), \quad (10)$$

since $\mathbb{E}[\zeta_t^{s+1}] = (1-\lambda)\nabla f(x_t^{s+1}) - \lambda\nabla f(\tilde{x}^s)$. The second inequality is obtain from Ineq. 9. And the last inequality, we followed the Eq. 2 and L-smooth function: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$.

Consider now the Lyapinov function:

$$R_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_t\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2]. \quad (11)$$

To bound Eq. 11, we require the bound of $\mathbb{E}[\|(1-\lambda)x_{t+1}^{s+1} - \lambda\tilde{x}^s\|^2]$ as following:

$$\begin{aligned}
&\mathbb{E}[\|(1-\lambda)x_{t+1}^{s+1} - \lambda\tilde{x}^s\|^2] \\
&= \mathbb{E}[\|(1-\lambda)(x_{t+1}^{s+1} - x_t^{s+1}) + (1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] \\
&= \mathbb{E}[\|(1-\lambda)x_{t+1}^{s+1} - x_t^{s+1}\|^2 + \|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2 + \\
&2\langle(1-\lambda)(x_{t+1}^{s+1} - x_t^{s+1}), ((1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s)\rangle] \\
&= \mathbb{E}[\eta_t^2(1-\lambda)^2\|v_t^{s+1}\|^2 + \|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] - \\
&2\eta_t(1-\lambda)\mathbb{E}[\langle\nabla f(x_t^{s+1}), (1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\rangle] \\
&\leq \mathbb{E}[(1-\lambda)^2\eta_t^2\|v_t^{s+1}\|^2 + \|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] + \\
&2\eta_t(1-\lambda)\mathbb{E}[\frac{1}{2\beta_t}\|\nabla f(x_t^{s+1})\|^2 + \frac{1}{2}\beta_t\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2]
\end{aligned} \quad (12)$$

The second equality follows from the unbiasedness of the update of Alg 2. The last inequality follows from application of Cauchy-Schwarz and Young's inequality. Combing Eq 9, Eq 11 and Eq 12, we can achieve the bound of $R_{t+1_{unbiased}}^{s+1} := \mathbb{E}[f(x_{t+1}^{s+1}) + c_{t+1}\|(1-\lambda)x_{t+1}^{s+1} - \lambda\tilde{x}^s\|^2]$ as

$$\begin{aligned}
&R_{t+1_{unbiased}}^{s+1} \leq \mathbb{E}[f(x_t^{s+1}) - \eta_t\|\nabla f(x_t^{s+1})\|^2 + \frac{L\eta_t^2}{2}\|v_t^{s+1}\|^2] + \\
&\mathbb{E}[c_{t+1}\eta_t^2(1-\lambda)^2\|v_t^{s+1}\|^2 + c_{t+1}\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] - \\
&2c_{t+1}(1-\lambda)\eta_t\mathbb{E}[\frac{1}{2\beta_t}\|\nabla f(x_t^{s+1})\|^2 + \frac{1}{2}\beta_t\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] \\
&\leq \mathbb{E}[f(x_t^{s+1}) - (\eta_t + \frac{c_{t+1}\eta_t(1-\lambda)}{\beta_t})\|\nabla f(x_t^{s+1})\|^2] + \\
&(\frac{L\eta_t^2}{2} + c_{t+1}\eta_t^2(1-\lambda)^2)\mathbb{E}[\|v_t^{s+1}\|^2] + \\
&(c_{t+1} + c_{t+1}\eta_t\beta_t(1-\lambda))\mathbb{E}[\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] \\
&= \mathbb{E}[f(x_t^{s+1})] - \\
&(\eta_t - \frac{c_{t+1}\eta t(1-\lambda)}{\beta_t} - (1-\lambda)^2L\eta_t^2 - 2(1-\lambda)^4c_{t+1}\eta_t^2)\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + \\
&\left(c_{t+1}(1+\eta_t\beta_t(1-\lambda)+2(1-\lambda)^2\eta^2L^2)+L^3\eta_t^2\right)\mathbb{E}[\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2] \\
&\leq R_t^{s+1} - (\eta_t - \frac{c_{t+1}\eta t(1-\lambda)}{\beta_t} - (1-\lambda)^2L\eta_t^2 - 2(1-\lambda)^4c_{t+1}\eta_t^2)\mathbb{E}[\|\nabla f(x_t^{s+1})
\end{aligned} \quad (13)$$

The last inequality follows $R_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_t\|(1-\lambda)x_t^{s+1} - \lambda\tilde{x}^s\|^2]$ where

$$c_{t_{unbiased}} = c_{t+1}(1+\eta_t\beta_t(1-\lambda)+2(1-\lambda)^2\eta_t^2L^2)+L^3\eta_t^2. \quad (14)$$

Thus the Ineq. 13 can be alternated as

$$\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{R_t^{s+1} - R_{t+1}^{s+1}}{\Omega_{t_{unbiased}}}, \quad (15)$$

where $\Omega_{t_{unbiased}} = \eta_t - \frac{c_{t+1}\eta_t(1-\lambda)}{\beta_t} - (1-\lambda)^2L\eta_t^2 - 2(1-\lambda)^4c_{t+1}\eta_t^2$

□

2 Proof of Theorem 5

Theorem. *Let* $f \in \mathcal{F}_n$, *let* $c_m = 0$, $\eta_t = \eta > 0$, $\beta_t = \beta > 0$, $c_{t_{unbiased}} = c_{t+1}(1+(1-\lambda)\eta\beta+2(1-\lambda)^2\eta^2L^2)+L^3\eta^2$, *so the* *intermediate result* $\Omega_{t_{\text{unbiased}}} = (\eta_t - (1-\lambda)\frac{c_{t+1}\eta_t}{\beta_t} - (1-\lambda)^2 L\eta_t^2 - 2(1-\lambda)^4 c_{t+1}\eta_t^2) > 0$*, for* $0 \le t \le m-1$*. Define the minimum value of* $\gamma_{n_{\text{unbiased}}} := \min_t \Omega_{t_{\text{unbiased}}}$*. Further let* $p_i = 0$ *for* $0 \le i < m$ *and* $p_m = 1$*, and* T *is a multiple of* m*. So the output* x_a *of Alg. 2 we have*

$$\mathbb{E}[\| \nabla f(x_a) \|^2] \le \frac{f(x^0) - f(x^*)}{T\gamma_{n_{\text{unbiased}}}},$$

where x^* *is the optimal solution to Problem 1.*

Proof. Using the result from Lemma 2 and $\eta_t = \eta$ when $t \in \{0, ..., m-1\}$, we can achieve the following bound:

$$\sum_{t=0}^{m-1} \mathbb{E}[\| \nabla f(x_t^{s+1}) \|^2] \le \frac{R_0^{s+1} - R_m^{s+1}}{\gamma_{n_{\text{unbiased}}}}, \tag{16}$$

Thus, the bound in Ineq. 16 can updated as

$$\sum_{t=0}^{m-1} \mathbb{E}[\| \nabla f(x_t^{s+1}) \|^2] \le \frac{\mathbb{E}[f(\tilde{x}^s) - f(\tilde{x}^{s+1})]}{\gamma_{n_{\text{unbiased}}}}, \tag{17}$$

where $R_0^{s+1} = \mathbb{E}[f(\tilde{x}^s)]$ since $x_0^{s+1} = \tilde{x}^s$ and $R_m^{s+1} = \mathbb{E}[f(\tilde{x}^{s+1})]$ since $x_m^{s+1} = \tilde{x}^{s+1}$, which we use the condition that $c_m = 0$, $p_m = 1$, and $p_i = 0$ for $i < m$. For the total number of iterations $T = Sm$, we further sum up iteration s as

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1} \mathbb{E}[\| \nabla f(x_t^{s+1}) \|^2] \le \frac{f(x^0) - f(x^*)}{T\gamma_{n_{\text{unbiased}}}}, \tag{18}$$

where the $\tilde{x}^0 = x^0$ and $\tilde{x}^* = x^*$. Thus, we can obtain our final result. □

3 Proof of Theorem 6

Theorem. *Suppose* $f \in \mathcal{F}_n$*, let* $\eta = \frac{1}{3Ln^{a\alpha}}$ *(*$0 \le a \le 1$*, and* $0 < \alpha \le 1$*),* $\beta = \frac{L}{n^{b\alpha}}$ *(*$b > 0$*),* $m_{\text{unbiased}} = \lfloor \frac{3n^{(3a+b)\alpha}}{(1-\lambda)} \rfloor$ *and* T *is the total number of iterations. Then, we can obtain the lower bound* $\gamma_{n_{\text{unbiased}}} \ge \frac{(1-\lambda)\nu}{9n^{(2a-b)\alpha}L}$ *in Theorem 5. For the output* x_a *of Alg. 2 we have*

$$\mathbb{E}[\| \nabla f(x_a) \|^2] \le \frac{9n^{(2a-b)\alpha}L[f(x^0) - f(x^*)]}{(1-\lambda)T\nu},$$

where x_* *is an optimal solution to Eq. 1.*

Proof. Using the relation in Eq 14 and $c_m = 0$, we estimated the upper bound of c_0 as

$$c_0 = L^3\eta^2 \frac{(1+\theta_{\text{unbiased}})^m - 1}{\theta_{\text{unbiased}}}, \tag{19}$$

where $\theta_{\text{unbiased}} = 2(1-\lambda)^2L^2\eta^2 + \eta\beta(1-\lambda)$. Let $\eta = \frac{1}{3Ln^{a\alpha}}$ and $\beta = \frac{L}{n^{b\alpha}}$, the θ can be alternated as:

$$\begin{aligned}\theta_{\text{unbiased}} &= 2(1-\lambda)^2L^2\eta^2 + \eta\beta(1-\lambda) = \frac{(1-\lambda)}{3n^{(a+b)\alpha}} + \frac{2(1-\lambda)^2}{9n^{2a\alpha}} \\ &\le \frac{1-\lambda}{3n^{(3a+b)\alpha}}.\end{aligned} \tag{20}$$

Using the above bound θ, we can get the further bound of c_0 as

$$\begin{aligned}c_0 &= \frac{L^3[(1+\theta_{\text{unbiased}})^m - 1]}{9L^2n^{2a\alpha}\left(\frac{1-\lambda}{3n^{(a+b)\alpha}} + \frac{2(1-\lambda)^2}{9n^{2a\alpha}}\right)} \\ &= \frac{L^3[(1+\theta_{\text{unbiased}})^m - 1]}{3L^2(1-\lambda)n^{(a-b)\alpha} + 2L^2(1-\lambda)^2} \le \frac{L^3(e-1)}{3L^2n^{(a-b)\alpha}},\end{aligned} \tag{21}$$

In the first inequality, due to the value of $(1+\theta_{\text{unbiased}})^{m_{\text{unbiased}}}$ is increasing when $m_{\text{unbiased}} = \lfloor \frac{1}{\theta_{\text{unbiased}}} \rfloor > 0$, we can use $\lim_{l\to\infty}(1+\frac{1}{l})^l = e$ (the e is Euler's number) to calculate upper bound of $(1+\theta)^{m_{\text{unbiased}}}$. Next, the lower bound of $\gamma_{n_{\text{unbiased}}}$ is given as:

$$\begin{aligned}\gamma_{n_{\text{unbiased}}} &= \min_t(\eta - \frac{c_{t+1}\eta}{\beta}(1-\lambda) - (1-\lambda)^2L\eta^2 - 2(1-\lambda)^4c_{t+1}\eta^2) \\ &\ge (\eta - \frac{c_0\eta}{\beta}(1-\lambda) - (1-\lambda)^2L\eta^2 - 2(1-\lambda)^4c_0\eta^2) \\ &\ge \frac{(1-\lambda)\nu}{9Ln^{(2a-b)\alpha}},\end{aligned} \tag{22}$$

where ν is independent of n. According to Theorem 5, we can achieve our result. □

4 Proof of theorem 7

Theorem. *Suppose* $f \in \mathcal{F}_n$ *have* σ*-bounded gradient. Let* $\eta_{t_{\text{biased}}} = \eta_\Delta = c_{\text{biased}}/\sqrt{\Delta+1}$ *for* $0 \le \Delta \le T-1$ *where* $c_{\text{biased}} = \sqrt{\frac{f(x_0) - f(x^*)}{2\lambda L\sigma^2}}$ *and let* T *be a multiple of* m*. Further let* $p_m = 1$*, and* $p_i = 0$ *for* $0 \le i < m$*. Then the output* x_a *of Alg. 3 we have*

$$\mathbb{E}[\| \nabla f(x_a)^2 \|] \le \frac{2(1-\lambda)}{\sqrt{\lambda}}\sqrt{\frac{2(f(x^0) - f(x^*))L}{T}}\sigma$$

Proof. As the learning rate decay from 1 to T, we use Definition 2 to bound gradients v_t^{s+1} as following:

$$\begin{aligned}&\mathbb{E}[\| v_t^{s+1} \|^2] \\ &= \mathbb{E}[\| (1-\lambda)(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)) + \lambda\nabla f(\tilde{x}^s) \|^2] \\ &= \mathbb{E}[\| (1-\lambda)\nabla f_{i_t}(x_t^{s+1}) - (1-\lambda)\nabla f_{i_t}(\tilde{x}^s) + \lambda\nabla f(\tilde{x}^s) \|^2] \\ &\le 2(\mathbb{E}[(\| (1-\lambda)\nabla f_{i_t}(x_t^{s+1}) \|^2 + \| (1-\lambda)\nabla f_{i_t}(\tilde{x}^s) - \lambda\nabla f(\tilde{x}^s) \|^2]) \\ &\le 2((1-\lambda)^2\mathbb{E}[\| \nabla f_{i_t}(x_t^{s+1}) \|^2] + (1-\lambda)^2\mathbb{E}[\| \nabla f_{i_t}(\tilde{x}^s) \|^2]) \\ &\le 4(1-\lambda)^2\sigma^2,\end{aligned} \tag{23}$$

where the first inequality we followed Lemma 3 when r=2. The second inequality we followed (a) σ-bounded gradient property of f and (b) the fact that for a random variable ζ which has a upper bounding as

$$\begin{aligned}&\mathbb{E}[\| (1-\lambda)\zeta - \lambda\mathbb{E}[\zeta] \|^2] \\ &= \mathbb{E}[(1-\lambda)^2 \| \zeta \|^2 - 2(1-\lambda)\lambda\zeta\mathbb{E}[\zeta] + \lambda^2\mathbb{E}^2[\zeta]] \\ &= (1-\lambda)^2\mathbb{E}[\| \zeta \|^2] - (2\lambda - 3\lambda^2)\mathbb{E}^2[\zeta] \\ &\le (1-\lambda)^2\mathbb{E}[\| \zeta \|^2],\end{aligned} \tag{24}$$

where the inequality should satisfy a condition that $0 \leq \lambda \leq \frac{2}{3}$.

Since f is $\mathcal{L}$-smooth, we have

$$\mathbb{E}[f(x_{t+1}^{s+1})] \leq \mathbb{E}[f(x_t^{s+1}) + \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1} \rangle + \frac{L}{2} \parallel x_{t+1}^{s+1} - x_t^{s+1} \parallel^2]. \tag{25}$$

Using Alg. 3 to update and since $\mathbb{E}[\langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1} \rangle] = \mathbb{E}[(\lambda - 2) \| \nabla f(x_t^{s+1}) \|^2]$ (unbiasedness of the stochastic gradients when $t \to \infty$), Ineq. 25 would be updated as:

$$\mathbb{E}[f(x_{t+1}^{s+1})] \leq \mathbb{E}[f(x_t^{s+1}) - \lambda\eta_\Delta \parallel \nabla f(x_t^{s+1}) \parallel^2 + \frac{L\eta_\Delta^2}{2} \parallel v_t^{s+1} \parallel^2]. \tag{26}$$

Adding the bound of v_t^{s+1} from Ineq. 23 to Ineq. 26, we can obtain that:

$$\mathbb{E}[f(x_{t+1}^{s+1})] \leq \mathbb{E}[f(x_t^{s+1})] - \lambda\eta_\Delta \mathbb{E}[\parallel \nabla f(x)_t^{s+1} \parallel^2] + \frac{L\eta_\Delta^2}{2}(4(1-\lambda)^2)\sigma^2 \tag{27}$$

Thus the Ineq. 27 can be alternated as

$$\mathbb{E}[\parallel \nabla f(x)_t^{s+1} \parallel^2] \leq \frac{1}{\eta_\Delta \lambda} \mathbb{E}[f(x_t^{s+1}) - f(x_{t+1}^{s+1})] + \frac{L\eta_\Delta}{\lambda}(2(1-\lambda)^2)\sigma^2, \tag{28}$$

where $t \in \{0, ..., m-1\}$, $s \in \{0, ..., S-1\}$, $\Delta \in \{0, ..., T-1\}$, and $T = mS$.

The minimum upper bound in Ineq. 29 can be achieved when $t = m-1$ and $s = S-1$, then we can obtain:

$$\begin{aligned} &\min_{t,s} \mathbb{E}[\parallel \nabla f(x_t^{s+1}) \parallel^2] \leq \\ &\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E}[\parallel f(x_t^{s+1}) \parallel^2] + \frac{L\eta_\Delta}{\lambda}(2(1-\lambda)^2)\sigma^2 \\ &\leq \frac{1}{T}\frac{1}{\eta\lambda} \mathbb{E}[f(x^0) - f(x^T)] + \frac{L\eta(2(1-\lambda)^2)}{\lambda}\sigma^2 \\ &\leq \frac{1}{T\eta\lambda}(f(x^0) - f(x^*)) + \frac{L\eta}{\lambda}(2(1-\lambda)^2)\sigma^2 \end{aligned} \tag{29}$$

The first inequality can hold due to the minimum is less than average. The second inequality is achieved from Eq 28, and the third one is followed the fact that $f(x^*) \leq f(x^T)$. To calculate learning rate $\eta_\Delta = \eta$, we take the derivative of the last inequality in Inequality 29 as

$$\frac{\partial \left(\frac{1}{T\eta\lambda}(f(x^0) - f(x^*)) + \frac{L\eta}{\lambda}(2(1-\lambda)^2)\sigma^2 \right)}{\partial \eta} = 0 \tag{30}$$

Thus, $\eta_\Delta = \eta = c/\sqrt{\Delta + 1}$, where $c = \sqrt{\frac{f(x^0) - f(x^*)}{2\lambda L \sigma^2}}$. Bring the result of $\eta_\Delta = \eta = c/\sqrt{\Delta + 1}$ to Eq. 29, we can achieve the upper bound of expectation as

$$\min_{t,s} \mathbb{E}[\parallel \nabla f(x_t^{s+1}) \parallel^2] \leq \frac{1}{\sqrt{T}}(\frac{1}{c\lambda}(f(x^0) - f(x^*)) + 2Lc\sigma^2). \tag{31}$$

□

For the case that the learning rate depends on the data size n, we provide one useful lemma in Lemma 2 firstly that can be used for proofing our Theorems.

Lemma 2. *For $c_t, c_{t+1}, \beta_t > 0$, we have*

$$c_{t_{biased}} = c_{t+1}(1 + \eta_t\beta_t + 2(1-\lambda)^2\eta_t^2 L^2) + L^3\eta_t^2(1-\lambda)^2.$$

Let η_t, β_t and c_{t+1} is given so that the $\Omega_t > 0$ can be showed as

$$\Omega_{t_{biased}} = \eta_t - \frac{c_{t+1}\eta_t}{\beta_t} - \lambda^2 L\eta_t^2 - 2\lambda^2 c_{t+1}\eta_t^2$$

Thus, the iterates in Alg. 3 satisfy the bound:

$$\mathbb{E}[\parallel \nabla f(x_t^{s+1}) \parallel^2] \leq \frac{R_{t_{biased}}^{s+1} - R_{t+1_{biased}}^{s+1}}{\Omega_{t_{biased}}}$$

where $R_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_{t_{biased}} \parallel x_t^{s+1} - \tilde{x}^s \parallel^2]$ for $0 \leq s \leq S-1$.

Proof. To further bound the result in Ineq. 26 since f is $\mathcal{L}$-smooth, we require to bound the intermediate iterates v_t^{s+1}, which is showed following inequalities:

$$\begin{aligned} &\mathbb{E}[\parallel v_t^{s+1} \parallel^2] \\ &= \mathbb{E}[\parallel (1-\lambda)(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)) + \lambda \nabla f(\tilde{x}^s)) \parallel^2] \\ &= \mathbb{E}[\parallel (1-\lambda)\zeta_t^{s+1} + \lambda \nabla f(\tilde{x}^s) - \lambda \nabla f(x_t^{s+1}) + \lambda \nabla f(x_t^{s+1}) \parallel^2] \\ &\leq 2\mathbb{E}[\parallel \lambda \nabla f(x_t^{s+1}) \parallel^2] + 2\mathbb{E}[\parallel (1-\lambda)\zeta_t^{s+1} - \lambda \mathbb{E}[\zeta_t^{s+1}] \parallel^2] \\ &\leq 2\lambda^2 \mathbb{E}[\parallel \nabla f(x_t^{s+1}) \parallel^2] + 2(1-\lambda)^2 \mathbb{E}[\parallel \nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s) \parallel^2] \\ &\leq 2\lambda^2 \mathbb{E}[\parallel \nabla f(x_t^{s+1}) \parallel^2] + 2(1-\lambda)^2 L^2 \mathbb{E}[\parallel x_t^{s+1} - \tilde{x}^s \parallel^2], \end{aligned} \tag{32}$$

where $0 \leq \lambda \leq 1$. In the first inequality, the variable ζ is showed as

$$\zeta_t^{s+1} = \frac{1}{|I_t|} \sum_{i_t \in I_t} (\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)), \tag{33}$$

since $\mathbb{E}[\zeta_t^{s+1}] = \nabla f(x_t^{s+1}) - \nabla f(\tilde{x}^s)$. The second inequality is obtain from Ineq. 24.

Consider now the Lyapinov function:

$$R_{t_{biased}}^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_{t_{biased}} \parallel x_t^{s+1} - \tilde{x}^s \parallel^2]. \tag{34}$$

To bound Eq. 34, we require the bound of $\mathbb{E}[\parallel x_{t+1}^{s+1} - \tilde{x}^s \parallel^2]$ as following:

$$\begin{aligned} &\mathbb{E}[\parallel x_{t+1}^{s+1} - \tilde{x}^s \parallel^2] \\ &= \mathbb{E}[\parallel x_{t+1}^{s+1} - x_t^{s+1} + x_t^{s+1} - \tilde{x}^s \parallel^2] \\ &= \mathbb{E}[\parallel x_{t+1}^{s+1} - x_t^{s+1} \parallel^2 + \parallel x_t^{s+1} - \tilde{x}^s \parallel^2 + 2\langle x_{t+1}^{s+1} - x_t^{s+1}, x_t^{s+1} - \tilde{x}^s \rangle] \\ &= \mathbb{E}[\eta_t^2 \parallel v_t^{s+1} \parallel^2 + \parallel x_t^{s+1} - \tilde{x}^s \parallel^2] - 2\eta_t \mathbb{E}[\langle \nabla f(x_t^{s+1}), x_t^{s+1} - \tilde{x}^s \rangle] \\ &\leq \mathbb{E}[\eta_t^2 \parallel v_t^{s+1} \parallel^2 + \parallel x_t^{s+1} - \tilde{x}^s \parallel^2] + \\ &2\eta_t \mathbb{E}[\frac{1}{2\beta_t} \parallel \nabla f(x_t^{s+1}) \parallel^2 + \frac{1}{2}\beta_t \parallel x_t^{s+1} - \tilde{x}^s \parallel^2] \end{aligned} \tag{35}$$

The second equality follows from the unbiasedness of the update of Alg 3. The last inequality follows from application of Cauchy-Schwarz and Young's inequality.

Combing Eq 32, Eq 34 and Eq 35, we can achieve the bound of $R^{s+1}_{t+1_{biased}} := \mathbb{E}[f(x^{s+1}_{t+1}) + c_{t+1} \parallel x^{s+1}_{t+1} - \tilde{x}^s \parallel^2]$ as

$$\begin{aligned}
&R^{s+1}_{t+1_{biased}} \leq \mathbb{E}[f(x^{s+1}_t) - \eta_t \parallel \nabla f(x^{s+1}_t) \parallel^2 + \frac{L\eta_t^2}{2} \parallel v^{s+1}_t \parallel^2] + \\
&\mathbb{E}[c_{t+1}\eta_t^2 \parallel v^{s+1}_t \parallel^2 + c_{t+1} \parallel x^{s+1}_t - \tilde{x}^s \parallel^2] + \\
&2c_{t+1}\eta_t \mathbb{E}[\frac{1}{2\beta_t} \parallel \nabla f(x^{s+1}_t) \parallel^2 + \frac{1}{2}\beta_t \parallel x^{s+1}_t - \tilde{x}^s \parallel^2] \\
&\leq \mathbb{E}[f(x^{s+1}_t) - (\eta_t - \frac{c_{t+1}\eta_t}{\beta_t}) \parallel \nabla f(x^{s+1}_t) \parallel^2] + \\
&(\frac{L\eta_t^2}{2} + c_{t+1}\eta_t^2)\mathbb{E}[\parallel v^{s+1}_t \parallel^2] + (c_{t+1} + c_{t+1}\eta_t\beta_t)\mathbb{E}[\parallel x^{s+1}_t - \tilde{x}^s \parallel^2] \\
&= \mathbb{E}[f(x^{s+1}_t)] - (\eta_t - \frac{c_{t+1}\eta t}{\beta_t} - \lambda^2 L\eta_t^2 - 2\lambda^2 c_{t+1}\eta_t^2)\mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] + \\
&\left(c_{t+1}(1 + \eta_t\beta_t + 2(1-\lambda)^2\eta^2L^2) + (1-\lambda)^2L^3\eta_t^2\right) \mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] \\
&\leq R^{s+1}_{t_{biased}} - (\eta_t - \frac{c_{t+1}\eta t}{\beta_t} - \lambda^2 L\eta_t^2 - 2\lambda^2 c_{t+1}\eta_t^2)\mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2].
\end{aligned} \tag{36}$$

The last inequality follows $R^{s+1}_{t_{biased}} := \mathbb{E}[f(x^{s+1}_t) + c_{t_{biased}} \parallel x^{s+1}_t - \tilde{x}^s \parallel^2]$ where

$$c_{t_{biased}} = c_{t+1}(1 + \eta_t\beta_t + 2(1-\lambda)^2\eta_t^2L^2) + (1-\lambda)^2L^3\eta_t^2. \tag{37}$$

Thus the Ineq. 36 can be alternated as

$$\mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] \leq \frac{R^{s+1}_{t_{biased}} - R^{s+1}_{t+1_{biased}}}{\Omega_{t_{biased}}}, \tag{38}$$

where $\Omega_{t_{biased}} = \eta_t - \frac{c_{t+1}\eta t}{\beta_t} - \lambda^2 L\eta_t^2 - 2\lambda^2 c_{t+1}\eta_t^2$ □

5 Proof of Theorem 8

Theorem. *Let* $f \in \mathcal{F}_n$, *let* $c_m = 0$, $\eta_t = \eta > 0$, $\beta_t = \beta > 0$, $c_{t_{biased}} = c_{t+1}(1 + \eta\beta + 2(1-\lambda)^2\eta^2L^2) + L^3\eta^2(1-\lambda)^2$, *so the intermediate result* $\Omega_{t_{biased}} = (\eta_t - \frac{c_{t+1}\eta_t}{\beta_t} - \lambda^2 L\eta_t^2 - 2\lambda^2 c_{t+1}\eta_t^2) > 0$, *for* $0 \leq t \leq m-1$. *Define the minimum value of* $\gamma_{n_{biased}} := \min_t \Omega_{t_{biased}}$. *Further let* $p_i = 0$ *for* $0 \leq i < m$ *and* $p_m = 1$, *and* T *is a multiple of* m. *So the output* x_a *of Alg. 3 we have*

$$\mathbb{E}[\parallel \nabla f(x_a) \parallel^2] \leq \frac{f(x^0) - f(x^*)}{T\gamma_{n_{biased}}},$$

where x^* *is the optimal solution to Problem 1.*

Proof. Using the result from Lemma 2 and $\eta_t = \eta$ when $t \in \{0, ..., m-1\}$, we can achieve the following bound:

$$\sum_{t=0}^{m-1} \mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] \leq \frac{R^{s+1}_0 - R^{s+1}_m}{\gamma_{n_{biased}}}, \tag{39}$$

Thus, the bound in Ineq. 39 can updated as

$$\sum_{t=0}^{m-1} \mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] \leq \frac{\mathbb{E}[f(\tilde{x}^s) - f(\tilde{x}^{s+1})]}{\gamma_{n_{biased}}}, \tag{40}$$

where $R^{s+1}_0 = \mathbb{E}[f(\tilde{x}^s)]$ since $x^{s+1}_0 = \tilde{x}^s$ and $R^{s+1}_m = \mathbb{E}[f(\tilde{x}^{s+1})]$ since $x^{s+1}_m = \tilde{x}^{s+1}$, which we use the condition that $c_m = 0$, $p_m = 1$, and $p_i = 0$ for $i < m$. For the total number of iterations $T = Sm$, we further sum up iteration s as

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1} \mathbb{E}[\parallel \nabla f(x^{s+1}_t) \parallel^2] \leq \frac{f(x^0) - f(x^*)}{T\gamma_{n_{biased}}}, \tag{41}$$

where the $\tilde{x}^0 = x^0$ and $\tilde{x}^* = x^*$. Thus, we can obtain our final result. □

6 Proof of Theorem 9

Theorem. *Suppose* $f \in \mathcal{F}_n$, *let* $\eta = \frac{1}{3Ln^{a\alpha}}$ *(*$0 \leq a \leq 1$ *and* $0 < \alpha \leq 1$*),* $\beta = \frac{L}{n^{b\alpha}}$ *(*$b > 0$*),* $m_{biased} = \lfloor \frac{3n^{2a\alpha}}{2(1-\lambda)} \rfloor$ *and* T *is the total number of iterations. Then, we can obtain the lower bound* $\gamma_{n_{biased}} \geq \frac{(1-\lambda)\lambda\nu_1}{9Ln^{(2a-b)\alpha}}$ *in Theorem 8. For the output* x_a *of Alg. 3 we have*

$$\mathbb{E}[\parallel \nabla f(x_a) \parallel^2] \leq \frac{9Ln^{(2a-b)\alpha}[f(x^0) - f(x^*)]}{\lambda(1-\lambda)T\nu_1},$$

where x_* *is an optimal solution to Eq. 1.*

Proof. Using the relation in Eq 37 and $c_m = 0$, we estimated the upper bound of c_0 as

$$c_0 = L^3\eta^2(1-\lambda)^2\frac{(1 + \theta_{biased})^m - 1}{\theta_{biased}}, \tag{42}$$

where $\theta_{biased} = 2(1-\lambda)^2L^2\eta^2 + \eta\beta$. Let $\eta = \frac{1}{3Ln^{a\alpha}}$ and $\beta = \frac{L}{n^{b\alpha}}$, the θ_{biased} can be alternated as:

$$\begin{aligned}
\theta_{biased} &= 2(1-\lambda)^2L^2\eta^2 + \eta\beta = \frac{2(1-\lambda)^2}{9n^{2a\alpha}} + \frac{1}{3n^{(a+b)\alpha}} \\
&\leq \frac{2(1-\lambda)}{3n^{2a\alpha}}.
\end{aligned} \tag{43}$$

Using the above bound θ, we can get the further bound of c_0 as

$$c_0 = \frac{(1-\lambda)^2L[(1 + \theta_{biased})^m - 1]}{2(1-\lambda)^2 + \frac{3}{n^{(b-a)\alpha}}} \leq \frac{L(1-\lambda)^2(e-1)}{3n^{(a-b)\alpha}}, \tag{44}$$

where $0 \leq \mu_0 \leq 1$ and $n \geq 1$. In the first inequality, due to the value of $(1+\theta_{biased})^{m_{biased}}$ is increasing when $m_{biased} = \lfloor \frac{1}{\theta} \rfloor > 0$, we can use $\lim_{l\to\infty}(1 + \frac{1}{l})^l = e$ (the e is Euler's number) to calculate upper bound of $(1 + \theta_{biased})^{m_{biased}}$. Next, the lower bound of $\gamma_{n_{biased}}$ is given as:

$$\begin{aligned}
\gamma_{n_{biased}} &= \min_t(\eta - \frac{c_{t+1}\eta}{\beta} - \lambda^2L\eta^2 - 2\lambda^2c_{t+1}\eta_t^2) \\
&\geq (\eta - \frac{c_0\eta}{\beta} - \lambda^2L\eta^2 - 2\lambda^2c_0\eta^2) \\
&\geq \frac{(1-\lambda)\lambda\nu_1}{\lambda Ln^{(2a-2b)\alpha}},
\end{aligned} \tag{45}$$

where ν_1 is independent of n. According to Theorem 8, we can achieve our result. □

7 Proof of Corollary 1

Corollary. *Suppose* $f \in \mathcal{F}_n$*, the IFO complexity of Alg. 4 (with parameters from Theorem 10) achieves an* ε*-accurate solution that is* $\mathcal{O}(\min\{1/\varepsilon^2, n^{1/5}/\varepsilon\})$*, where the number of IFO calls is minimized when* $a = 1$*,* $b = 2$ *and* $\alpha = 1/5$*.*

Proof. This result of IFO is $\mathcal{O}(\min\{1/\varepsilon^2, n^{1/5}/\varepsilon\})$. For the first term of IFO follows from Theorem 7, it is same with SGD IFO calls.

For the second term of IFO follows from Theorem 6 and fact that $m = \lfloor \frac{3n^{(3a+b)\alpha}}{(1-\lambda)} \rfloor$. Suppose $\alpha < \frac{1}{(3a+b)}$, then $m = o(n)$. However, n IFO calls invested in calculating the average gradient at the end of each epoch. In other words, computation of average gradient requires n IFO calls for every m iterations of algorithm. Using this relationship, we get $O(n + n^{(1-\frac{\alpha}{2})\varepsilon})$ in this case. On the other hand, when $\alpha > \frac{1}{(3a+b)}$, the total number of IFO calls made by Alg 4 in each epoch is $\Omega(n)$ since $m = \lfloor \frac{3n^{(3a+b)\alpha}}{(1-\lambda)} \rfloor$. As a result, the oracle calls required for calculating the average gradient (per epoch) is of lower order, leading to $O(n + n^{\alpha}/\varepsilon)$ IFO calls. Consequently, $\alpha = \frac{1}{(3a+b)}$ is key result to achieve IFO calls as following:

To achieve a lowest upper bound in Theorem 10, the best choice is $a = 1$, $b = 2$. Thus, $\alpha = \frac{1}{5}$, and IFO in second case is $n^{1/5}/\varepsilon$. $\square$

Lemma 3. *For random variables* $z_1,\ldots,z_r$*, we have*

$$\mathbb{E}[\| z_1 + \ldots + z_r \|^2] \leq r\mathbb{E}[\| z_1 \|^2 + \ldots + \| z_r \|^2]. \quad (46)$$