APPENDIX A

Implementation Details 2:

In accordance with the fairness requirement, we have set algorithm-specific step size. The details are shown in the Table VI. The initial step size η_0 is found by a grid search from $\{0.01, 0.1, 1, 10, 100, 1000\}$, and the optimal η_0 is 1.

TABLE V
THE ALGORITHM-SPECIFIC STEP SIZE

Algorithms	Fed-ZO-GD	Fed-ZO-SGD	Fed-ZO-SignSGD	Fed-ZO-SVRG	DES	DZO-SNGM
Step-size	$\frac{\eta_0}{(t+1)^{0.5}(k+1)}$	$\frac{\eta_0}{(t+1)^{0.5}(k+1)^{0.5}}$	$\frac{\eta_0}{(t+1)^{0.5}}$	$\frac{\eta_0}{(t+1)^{0.5}(k+1)^{0.5}}$	$\frac{\eta_0}{(t+1)^{0.25}(k+1)^{0.5}}$	$\frac{\eta_0}{(t+1)^{0.75}}$

APPENDIX B

Proof of Lemma 2:

Proof: The first equality is well-known and the proof can be found in [16]. Note that the function h is convex in [16], but we can prove it without relying on the convexity. In contrast, the function h in this paper is assumed to be non-convex. The second result presented in Lemma 2 is proved as follows:

$$\mathbb{E}\left[\|g(\boldsymbol{x})\|^2\right] = \mathbb{E}\left[\left\|\frac{h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v})}{2\eta}\boldsymbol{v}\right\|^2\right] = \frac{1}{4\eta^2}\mathbb{E}\left[\|h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v})\|^2\|\boldsymbol{v}\|^2\right]. \tag{18}$$

According to the function h is L-smooth from Assumption 1, we have

$$|h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}) - \eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle| \leq \frac{L}{2}\eta^2\|\boldsymbol{v}\|^2, \tag{19}$$

$$|h(\boldsymbol{x}-\eta\boldsymbol{v}) - h(\boldsymbol{x}) + \eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle| \leq \frac{L}{2}\eta^2\|\boldsymbol{v}\|^2. \tag{20}$$

We can subtract (19) from (20),

$$-L\eta^2\|\boldsymbol{v}\|^2 \leq h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v}) - 2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle \leq L\eta^2\|\boldsymbol{v}\|^2, \tag{21}$$

and take the absolute value of (21),

$$|h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v}) - 2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle| \leq L\eta^2\|\boldsymbol{v}\|^2. \tag{22}$$

Then, we can push forward with deducing additional consequences from the second equality of (18),

$$\begin{aligned}
\mathbb{E}\left[\|g(\boldsymbol{x})\|^2\right] &= \frac{1}{4\eta^2}\mathbb{E}\left[\|h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v}) - 2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle + 2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle\|^2\|\boldsymbol{v}\|^2\right] \\
&\overset{(a)}{\leq} \frac{1}{2\eta^2}\mathbb{E}\left[\|h(\boldsymbol{x}+\eta\boldsymbol{v}) - h(\boldsymbol{x}-\eta\boldsymbol{v}) - 2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle\|^2\|\boldsymbol{v}\|^2\right] + \frac{1}{2\eta^2}\mathbb{E}\left[\|2\eta\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle\|^2\|\boldsymbol{v}\|^2\right] \\
&\overset{(b)}{\leq} \frac{1}{2\eta^2}\mathbb{E}\left[L^2\eta^4\|\boldsymbol{v}\|^6\right] + 2\mathbb{E}[\|\langle\nabla h(\boldsymbol{x}), \boldsymbol{v}\rangle\|^2] \\
&\overset{(c)}{\leq} \frac{L^2\eta^2}{2}\mathbb{E}[\|\boldsymbol{v}\|^6] + 2(n+4)\mathbb{E}[\|\nabla h(\boldsymbol{x})\|^2] \\
&\overset{(d)}{\leq} \frac{L^2\eta^2}{2}(n+6)^3 + 2(n+4)\|\nabla h(\boldsymbol{x})\|^2 \\
&\overset{(e)}{\leq} \frac{L^2\eta^2}{2}(n+6)^3 + 10n\|\nabla h(\boldsymbol{x})\|^2.
\end{aligned}$$

where (a) is due to the Jensen's inequality, (b) is due to (22). In consequence of the properties proposed in [18], we have the inequalities of (c) and (d). (e) is due to the fact that $2(n+4) \leq 10n$. ■

Proof of Lemma 3:

Proof: We can use the second inequality in Lemma 2 to $\nabla F(\boldsymbol{x}_{i,k}^t; \boldsymbol{\xi}_{i,k}^t)$,

$$\begin{aligned}
\mathbb{E}\left[\|g_{i,k}^t\|^2\right] &\leq \frac{L^2\eta^2}{2}(n+6)^3 + 10n\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_{i,k}^t; \boldsymbol{\xi}_{i,k}^t)\|^2\right] \\
&\overset{(a)}{\leq} \frac{L^2\eta^2}{2}(n+6)^3 + 10n(\|f(\boldsymbol{x})\|^2 + \sigma_i^2) \\
&\overset{(b)}{\leq} \frac{L^2\eta^2}{2}(n+6)^3 + 10n(G^2 + \sigma_i^2),
\end{aligned}$$

where (a) follows from Lemma 2 and (b) is due to the Assumption 3. ■

Proof of Lemma 4:

Proof: By the definition of $\boldsymbol{e}_t$, we have:

$$\begin{aligned}
\boldsymbol{e}_{t+1} &= \boldsymbol{\nu}_{t+1} + \eta K \nabla f_\eta(\boldsymbol{x}_{t+1}) \\
&\overset{(a)}{=} (1-\beta)\boldsymbol{\nu}_t + \beta\Delta_{t+1} + \eta K \nabla f_\eta\big(\boldsymbol{x}_{t+1}\big) \\
&\overset{(b)}{=} (1-\beta)\boldsymbol{e}_t - (1-\beta)\eta K \nabla f_\eta\big(\boldsymbol{x}_t\big) + \beta\Delta_{t+1} + \eta K\big((1-\beta)+\beta\big)\nabla f_\eta\big(x_{t+1}\big) \\
&= (1-\beta)\boldsymbol{e}_t - (1-\beta)\underbrace{\eta K\left(\nabla f_\eta(\boldsymbol{x}_t) - \nabla f_\eta(\boldsymbol{x}_{t+1})\right)}_{\boldsymbol{\psi}_{t+1}} + \underbrace{\beta\left(\Delta_{t+1} + \eta K \nabla f_\eta(\boldsymbol{x}_{t+1})\right)}_{\boldsymbol{\phi}_{t+1}} \\
&= (1-\beta)\boldsymbol{e}_t - (1-\beta)\eta K \boldsymbol{\psi}_{t+1} + \beta\boldsymbol{\phi}_{t+1} \\
&\overset{(c)}{=} (1-\beta)^{t+1}\boldsymbol{e}_0 - (1-\beta)\eta K \sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\psi}_\tau + \beta\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau,
\end{aligned}$$

where (a) and (b) follow from the definition of $\boldsymbol{\nu}_t$. We can derive (c) from unrolling the recursion of $\boldsymbol{e}_t$ over t rounds. We then take the norm of both sides of (b),

$$\|\boldsymbol{e}_{t+1}\| = (1-\beta)^{t+1}\|\boldsymbol{e}_0\| + (1-\beta)\eta K\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\psi}_\tau\right\| + \beta\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|, \tag{23}$$

where $\boldsymbol{\psi}_\tau$ accounts for the change of ∇f_η at τ-iteration. We know that ∇f_η is L-smooth from Lemma 1. Then,

$$\begin{aligned}
(1-\beta)\eta K\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\psi}_\tau\right\| &\overset{(a)}{\leq} (1-\beta)\eta K\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\|\boldsymbol{\psi}_\tau\| \\
&= (1-\beta)\eta K\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\|\nabla f_\eta(\boldsymbol{x}_{\tau-1} - \nabla f_\eta(\boldsymbol{x}_\tau)\| \\
&\overset{(b)}{\leq} (1-\beta)\eta K\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}L\|\boldsymbol{x}_{\tau-1} - \boldsymbol{x}_\tau\| \\
&= (1-\beta)\eta K\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}L\|\eta\frac{\boldsymbol{\nu}_\tau}{\|\boldsymbol{\nu}_\tau\|}\| \\
&= (1-\beta)\eta K\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}L\eta \\
&\overset{(c)}{\leq} \frac{KL\eta^2}{\beta},
\end{aligned}$$

where (a) follows from Jensens inequality, (b) is because of Lemma 1. (c) is due to the fact that $\frac{1-\beta}{\beta} \leq \frac{1}{\beta}, \beta \in [0,1]$. Then, (23) can be rewritten as:

$$\|\boldsymbol{e}_{t+1}\| = (1-\beta)^{t+1}\|\boldsymbol{e}_0\| + \frac{KL\eta^2}{\beta} + \beta\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|, \tag{24}$$

Taking total expectation gives that:

$$\mathbb{E}[\|\boldsymbol{e}_{t+1}\|] = (1-\beta)^{t+1}\mathbb{E}[\|\boldsymbol{e}_0\|] + \beta\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|\right] + \frac{KL\eta^2}{\beta}. \tag{25}$$

■

Proof of Lemma 5:

Proof: By the definition of $\boldsymbol{\phi}_\tau$:

$$\begin{aligned}
\boldsymbol{\phi}_\tau &= \Delta_\tau + \eta K \nabla f_\eta(x_\tau) \\
&= \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{x}_{i,K}^\tau - \boldsymbol{x}_\tau + \eta K \nabla f_\eta(x_\tau) \\
&= \frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} (\boldsymbol{x}_{i,k+1}^\tau - \boldsymbol{x}_{i,k}^\tau) + \eta K \nabla f_\eta(\boldsymbol{x}_\tau) \\
&= \frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \big(\nabla f_\eta(\boldsymbol{x}_\tau) - g_{i,k}^\tau\big) \\
&= \frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \big(\nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau) - g_{i,k}^\tau - \nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau) + \nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau) - \nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau) + \nabla f_\eta(\boldsymbol{x}_\tau)\big) \\
&= \frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \underbrace{\big(\nabla f_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau) - g_{i,k}^\tau\big)}_{\mathfrak{A}^\tau} + \frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \underbrace{\big(\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau) - \nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau)\big)}_{\mathfrak{B}^\tau} \\
&\quad + \frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \underbrace{\big(\nabla f_\eta(\boldsymbol{x}_\tau) - \nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)\big)}_{\mathfrak{C}^\tau},
\end{aligned}$$

where $\mathfrak{A}^\tau$ is the variance of gradient estimation, $\mathfrak{B}^\tau$ is the client-specific drift, and $\mathfrak{C}^\tau$ is the local sampling noise. By the unbiasedness property given in Lemma 2, we know that $\mathbb{E}[g_{i,k}^\tau] = \nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)$, which implies that $\mathbb{E}[\mathfrak{A}^\tau] = 0$. On the other hand, since the client-side sampling is unbiased, we have:

$$\mathbb{E}_{\boldsymbol{\xi}_i}\left[\frac{1}{M}\sum_{i=1}^{M}\nabla F(\boldsymbol{x};\boldsymbol{\xi}_i)\right] = \nabla f(\boldsymbol{x}), \tag{26}$$

which implies that

$$\mathbb{E}_{\boldsymbol{\nu}}\left[\mathbb{E}_{\boldsymbol{\xi}_i}\left[\frac{1}{M}\sum_{i=1}^{M}\nabla F(\boldsymbol{x};\boldsymbol{\xi}_i)\right]\right] = \mathbb{E}_{\boldsymbol{\nu}}\left[\nabla f(\boldsymbol{x}+\eta\boldsymbol{\nu})\right]. \tag{27}$$

Then, we can easily have

$$\mathbb{E}_{\boldsymbol{\xi}_i}\left[\frac{1}{M}\sum_{i=1}^{M}\nabla F_\eta(\boldsymbol{x};\boldsymbol{\xi}_i)\right] = \nabla f_\eta(\boldsymbol{x}), \tag{28}$$

So, the expectation of the mean of zeroth-order gradient estimates for all local clients is the global zeroth-order gradient estimates, i.e., $\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)\right] = \nabla f_\eta(\boldsymbol{x}_\tau)$ and $\mathbb{E}[\mathfrak{C}^\tau] = 0$. The variances of $\mathfrak{A}^\tau$ and $\mathfrak{C}^\tau$ will be diminished by accumulating $\boldsymbol{\phi}_\tau$ since their expectation are unbiased. We then have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{A}^\tau + \sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{B}^\tau + \sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{C}^\tau\right\|^2\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{A}^\tau\right\|^2\right] + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{B}^\tau\right\|^2\right] + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{C}^\tau\right\|^2\right] \\
&\overset{(b)}{=} \sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{A}^\tau\|^2\right] + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{B}^\tau\right\|^2\right] + \sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{C}^\tau\|^2\right],
\end{aligned}$$

where (a) is the due to $\mathbb{E}[\mathfrak{A}^\tau] = 0$ and $\mathbb{E}[\mathfrak{C}^\tau] = 0$. (b) follows from the Jensen's inequality. where the second term of the right

side of the last equality can be written as,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{B}^\tau\right\|^2\right] &= \left(\frac{1-(1-\beta)^{t+1}}{\beta}\right)^2\mathbb{E}\left[\left\|\frac{\beta}{1-(1-\beta)^{t+1}}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathfrak{B}^\tau\right\|^2\right]\\ &\overset{(a)}{\leq}\left(\frac{1-(1-\beta)^{t+1}}{\beta}\right)^2\frac{\beta}{1-(1-\beta)^{t+1}}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right]\\ &= \frac{1-(1-\beta)^{t+1}}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right]\\ &\overset{(b)}{\leq}\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right].\end{aligned}$$

where (a) is due to Jensen's inequality. (b) is due to $\beta \leq 1$. Then, we have

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|^2\right] \leq \sum_{\tau=1}^{t+1}(1-\beta)^{2(t+1-\tau)}\mathbb{E}\left[\|\mathfrak{A}^\tau\|^2+\|\mathfrak{C}^\tau\|^2\right]+\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right]. \tag{29}$$

where

$$\begin{aligned}\mathbb{E}\left[\|\mathfrak{A}^\tau\|^2\right] &= \frac{\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\mathbb{E}\left[\|g_{i,k}^\tau-\nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau)\|^2\right]\\ &\overset{(a)}{\leq}\frac{\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\mathbb{E}\left[g_{i,k}^\tau\right] \overset{(b)}{\leq}\frac{K\eta^2}{M}\sum_{i=1}^{M}\Phi_{i,\eta}=K\eta^2\Phi_\eta,\end{aligned}$$

where we use $\mathbb{E}[g_{i,k}^\tau]=\nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau)$ following from Lemma 2 to obtain (a). (b) is due to the fact that $\mathbb{E}[\|a-\mathbb{E}[a]\|^2]\leq\mathbb{E}[\|a\|^2]$.

$$\mathbb{E}\left[\|\mathfrak{C}^\tau\|^2\right]=\frac{\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)-\nabla f_\eta(\boldsymbol{x}_\tau)\|^2\right]\overset{(a)}{\leq}\frac{\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\sigma_i^2\overset{(b)}{=}K\eta^2\sigma^2, \tag{30}$$

where (a) and (b) are derived by $\|\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)-\nabla f_\eta(\boldsymbol{x}_\tau)\|^2 \leq \mathbb{E}_{\boldsymbol{\nu}}\left[\|\nabla F(\boldsymbol{x}_\tau+\eta\boldsymbol{\nu};\boldsymbol{\xi}_{i,k}^\tau)-\nabla f(\boldsymbol{x}_\tau+\eta\boldsymbol{\nu})\|^2\right]\leq\sigma_i^2$. Therefore, (29) can be rewritten as

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|^2\right] \leq K\eta^2\sum_{\tau=1}^{t+1}(1-\beta)^{2(t+1-\tau)}(\Phi_\eta+\sigma^2)+\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right] \tag{31a}$$

$$\leq \frac{K\eta^2(\Phi_\eta+\sigma^2)}{1-(1-\beta)^2)}+\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right] \tag{31b}$$

$$\leq \frac{K\eta^2(\Phi_\eta+\sigma^2)}{\beta}+\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right], \tag{31c}$$

where we can use $\beta\leq 1$ and $1-(1-\beta)^2=\beta(2-\beta)\geq\beta$ to obtain (31c) and (31) respectively. Next, we need to bound the last term of (31). Let $\Phi_{i,\eta}:=\frac{L^2\eta^2}{2}(n+6)^3+10n(G^2+\sigma_i^2)$ and $\Phi_\eta:=\sum_{i=1}^{M}\Phi_{i,\eta}$.

$$\begin{aligned}\mathbb{E}\left[\|\mathfrak{B}^\tau\|^2\right] &\leq \mathbb{E}\left[\left\|\frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau)-\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)\right\|^2\right]\\ &\overset{(a)}{\leq}\frac{K\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla F_\eta(\boldsymbol{x}_{i,k}^\tau;\boldsymbol{\xi}_{i,k}^\tau)-\nabla F_\eta(\boldsymbol{x}_\tau;\boldsymbol{\xi}_{i,k}^\tau)\right\|^2\right]\\ &\overset{(b)}{\leq}\frac{K\eta^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}L^2\mathbb{E}\left[\left\|\boldsymbol{x}_{i,k}^\tau-\boldsymbol{x}_\tau\right\|^2\right]\\ &= \frac{K\eta^2L^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\sum_{l=0}^{k-1}\mathbb{E}\left[\left\|\boldsymbol{x}_{i,l+1}^\tau-\boldsymbol{x}_{i,l}^\tau\right\|^2\right]\\ &\leq \frac{K^2\eta^2L^2}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\sum_{l=0}^{k-1}\mathbb{E}\left[\left\|\eta g_{i,l+1}^\tau\right\|^2\right]\overset{(c)}{\leq}\frac{K^2L^2\eta^4}{M}\sum_{i=1}^{M}\sum_{k=0}^{K}\sum_{l=0}^{k-1}\Phi_{i,\eta}=K^4L^2\eta^4\Phi_\eta,\end{aligned}$$

where (a) is due to Jensen's inequality, and (b) follows from Lemma 1, and (c) is because of Lemma 3. We can bound (29) by substituting the above inequality into (31),

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|^2\right] &\leq \frac{K\eta^2(\Phi_\eta+\sigma^2)}{\beta}+\frac{1}{\beta}\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}K^4L^2\eta^4\Phi_\eta \\ &\leq \frac{K\eta^2(\Phi_\eta+\sigma^2)}{\beta}+\frac{K^2L^2\eta^4\Phi_\eta}{\beta^2},\end{aligned}$$

which implies that

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|\right] \leq \sqrt{\frac{K\eta^2(\Phi_\eta+\sigma^2)}{\beta}+\frac{K^4L^2\eta^4\Phi_\eta}{\beta^2}}. \tag{32}$$

■

Proof of Lemma 6:

Proof: By the L-smooth property of f_η from Assumption 1, we have

$$f_\eta(\boldsymbol{x}_{t+1}) \leq f_\eta(\boldsymbol{x}_t)+\langle\nabla f_\eta(\boldsymbol{x}_t),\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle+\frac{L}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2 \tag{33a}$$

$$= f_\eta(\boldsymbol{x}_t)-\eta\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle+\frac{L}{2}\eta^2. \tag{33b}$$

Now we need to discuss it in two cases: 1) $\|K\eta\nabla f_\eta(\boldsymbol{x}_t)\|\geq 2\|\boldsymbol{e}_t\|$; 2) $\|K\eta\nabla f_\eta(\boldsymbol{x}_t)\|<2\|\boldsymbol{e}_t\|$. For case 1, we have

$$\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle=\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{e}_t+K\eta\nabla f_\eta(\boldsymbol{x}_t)}{\|\boldsymbol{e}_t+K\eta\nabla f_\eta(\boldsymbol{x}_t)\|}\right\rangle=\frac{\langle\nabla f_\eta(\boldsymbol{x}_t),\boldsymbol{e}_t\rangle}{\|\boldsymbol{e}_t+K\eta\nabla f_\eta(\boldsymbol{x}_t)\|}+\frac{K\eta\|\nabla f_\eta(\boldsymbol{x}_t)\|^2}{\|\boldsymbol{e}_t+K\eta\nabla f_\eta(\boldsymbol{x}_t)\|}, \tag{34}$$

where

$$\langle\nabla f_\eta(\boldsymbol{x}_t),\boldsymbol{e}_t\rangle\geq\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{-K\eta\nabla f_\eta(\boldsymbol{x}_t)}{2}\right\rangle=-\frac{K\eta\|\nabla f_\eta(\boldsymbol{x}_t)\|^2}{2}, \tag{35}$$

Substitute it into (34) ,

$$\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle\geq\frac{K\eta\|\nabla f_\eta(\boldsymbol{x}_t)\|^2}{2\|\boldsymbol{e}_t+K\eta\nabla f_\eta(\boldsymbol{x}_t)\|}, \tag{36}$$

where the denominator can be maximized by choosing $\boldsymbol{e}_t:=\frac{K\eta\nabla f_\eta(\boldsymbol{x}_t)}{2}$, that is,

$$\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle\geq\frac{K\eta\|\nabla f_\eta(\boldsymbol{x}_t)\|^2}{2\|\frac{3}{2}K\eta\nabla f_\eta(\boldsymbol{x}_t)\|}=\frac{\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}\geq\frac{\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}-\frac{8\|\boldsymbol{e}_t\|}{3K\eta}. \tag{37}$$

For case 2, we have

$$\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle\geq-\|\nabla f_\eta(\boldsymbol{x}_t)\|=\frac{\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}-\frac{4\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}\geq\frac{\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}-\frac{8\|\boldsymbol{e}_t\|}{3K\eta}. \tag{38}$$

Combing (37) and (38), we have

$$f_\eta(\boldsymbol{x}_{t+1})\leq f_\eta(\boldsymbol{x}_t)-\eta\left\langle\nabla f_\eta(\boldsymbol{x}_t),\frac{\boldsymbol{\nu}_t}{\|\boldsymbol{\nu}_t\|}\right\rangle+\frac{L\eta^2}{2}\leq f_\eta(\boldsymbol{x}_t)-\frac{\eta\|\nabla f_\eta(\boldsymbol{x}_t)\|}{3}+\frac{8\|\boldsymbol{e}_t\|}{3K}+\frac{L\eta^2}{2}. \tag{39}$$

■

Proof of Theorem 1:

Proof: By Lemmas 4 and 5,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}_{t+1}\|] &\leq (1-\beta)^{t+1}\mathbb{E}[\|\boldsymbol{e}_0\|]+\frac{KL\eta^2}{\beta}+\beta\mathbb{E}\left[\left\|\sum_{\tau=1}^{t+1}(1-\beta)^{t+1-\tau}\boldsymbol{\phi}_\tau\right\|\right] \\ &\leq (1-\beta)^{t+1}\mathbb{E}[\|\boldsymbol{e}_0\|]+\frac{KL\eta^2}{\beta}+\beta\sqrt{\frac{K\eta^2(\Phi_\eta+\sigma^2)}{\beta}+\frac{K^4L^2\eta^4\Phi_\eta}{\beta^2}} \\ &= (1-\beta)^{t+1}\mathbb{E}[\|\boldsymbol{e}_0\|]+\frac{KL\eta^2}{\beta}+\eta\sqrt{\beta K}\sqrt{\Phi_\eta+\sigma^2+\frac{K^3L^2\Phi_\eta\eta^2}{\beta}},\end{aligned}$$

Average over $t = 0, \ldots, T-1$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{e}_t\|] \leq \frac{1}{T}\sum_{t=0}^{T-1}(1-\beta)^t\mathbb{E}[\|\boldsymbol{e}_0\|] + \frac{KL\eta}{\beta} + \eta\sqrt{\beta K}\sqrt{\Phi_\eta + \sigma^2 + \frac{K^3L^2\Phi_\eta\eta^2}{\beta}} \tag{40a}$$

$$= \frac{\mathbb{E}[\|\boldsymbol{e}_0\|]}{T\beta} + \frac{KL\eta^2}{\beta} + \eta\sqrt{\beta K}\sqrt{\Phi_\eta + \sigma^2 + \frac{K^3L^2\Phi_\eta\eta^2}{\beta}}. \tag{40b}$$

By Lemma 6,

$$\mathbb{E}[f_\eta(\boldsymbol{x}_{t+1}) - f_\eta(\boldsymbol{x}_t)] \leq -\frac{\eta\mathbb{E}[\|\nabla f_\eta(\boldsymbol{x}_t)\|]}{3} + \frac{8\mathbb{E}[\|\boldsymbol{e}_t\|]}{3K} + \frac{L\eta^2}{2}. \tag{41}$$

Average over $t = 0, \ldots, T-1$:

$$\frac{1}{T}\mathbb{E}[f_\eta(\boldsymbol{x}_T) - f_\eta(\boldsymbol{x}_0)] \leq -\frac{\eta}{3T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\eta(\boldsymbol{x}_t)\|] + \frac{8}{3KT}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{e}_t\|] + \frac{L\eta^2}{2}. \tag{42}$$

By rearranging (42), we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\eta(\boldsymbol{x}_t)\|] \leq \frac{3}{T\eta}\mathbb{E}[f_\eta(\boldsymbol{x}_0) - f_\eta(\boldsymbol{x}_T)] + \frac{8}{KT\eta}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{e}_t\|] + \frac{3L\eta^2}{2} \tag{43a}$$

$$\leq \frac{3}{T\eta}\mathbb{E}[f_\eta(\boldsymbol{x}_0) - f_\eta(\boldsymbol{x}_T) + nL\eta^2] + \frac{8}{\eta KT}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{e}_t\|] + \frac{3L\eta^2}{2} \tag{43b}$$

$$\leq \frac{3}{T\eta}\left(\Delta_f + nL\eta^2\right) + \frac{8}{\eta KT}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{e}_t\|] + \frac{3L\eta^2}{2} \tag{43c}$$

$$\leq \frac{3}{T\eta}\left(\Delta_f + nL\eta^2\right) + \frac{8}{\eta K}\left(\frac{\mathbb{E}[\|\boldsymbol{e}_0\|]}{T\beta} + \frac{KL\eta^2}{\beta} + \eta\sqrt{\beta K}\sqrt{\Phi_\eta + \sigma^2 + \frac{K^3L^2\Phi_\eta\eta^2}{\beta}}\right) + \frac{3L\eta^2}{2}, \tag{43d}$$

where we can substitute (40b) into (43c) to obtain 43. By setting $\eta := \frac{\eta_0}{T^{3/4}} = \eta_0\beta^{\frac{3}{2}}$ and substituting them into (43d):

$$\frac{3}{T}\left(\frac{\Delta_f}{\eta_0\beta^{\frac{3}{2}}} + Ln\eta_0\beta^{\frac{3}{2}}\right) + \frac{8}{K}\left(\frac{\mathbb{E}[\|\boldsymbol{e}_0\|]}{T^{1/4}\beta\eta_0} + KL\eta_0\sqrt{\beta} + \sqrt{\beta K}\sqrt{\Phi_\eta + \sigma^2 + K^3L^2\Phi_\eta\eta_0^2\beta^2}\right) + \frac{3L\eta_0\beta^{\frac{3}{2}}}{2}. \tag{44}$$

If we choose $\beta := \frac{1}{\sqrt{T}}$, (44) can be rewritten as,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\eta(\boldsymbol{x}_t)\|] \leq \frac{3}{T}\left(\frac{\Delta_f T^{\frac{3}{4}}}{\eta_0} + nL\eta_0 T^{-\frac{3}{4}}\right) + T^{\frac{1}{4}}\frac{8\mathbb{E}[\|\boldsymbol{e}_0\|]}{K\eta_0} + T^{-\frac{1}{4}}\left(8L\eta_0 + \sqrt{\frac{\Phi_\eta + \sigma^2}{K} + K^2L^2\Phi_\eta\eta_0^2}\right) + \frac{3L\eta_0}{2T^{\frac{3}{4}}}$$

$$= T^{-\frac{1}{4}}\left(\frac{3\Delta_f}{\eta_0} + 8L\eta_0 + \sqrt{\frac{\Phi_\eta + \sigma^2}{K} + K^2L^2\Phi_\eta\eta_0^2}\right) + T^{\frac{1}{4}}\frac{8\mathbb{E}[\|\boldsymbol{e}_0\|]}{K\eta_0} + T^{-\frac{3}{4}}\frac{3L\eta_0}{2} + T^{-\frac{7}{4}}3nL\eta_0.$$

By Lemma 1, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\eta(\boldsymbol{x}_t)\|] + \frac{\eta L(n+3)^{\frac{3}{2}}}{2} \tag{45a}$$

$$\leq T^{-\frac{1}{4}}\left(\frac{3\Delta_f}{\eta_0} + 8L\eta_0 + \sqrt{\frac{\Phi_\eta + \sigma^2}{K} + K^2L^2\Phi_\eta\eta_0^2}\right) + T^{\frac{1}{4}}\frac{8\mathbb{E}[\|\boldsymbol{e}_0\|]}{K\eta_0} + T^{-\frac{3}{4}}(n+3)^{\frac{3}{2}}L\eta_0 + T^{-\frac{7}{4}}3nL\eta_0. \tag{45b}$$

Now, we only need to bound $\mathbb{E}[\|\boldsymbol{e}_0\|]$,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{e}_0\|] &= \mathbb{E}[\|\boldsymbol{\Delta}_0 + \eta K \nabla f_\eta(\boldsymbol{x}_0)\|] \\
&\overset{(a)}{\leq} \mathbb{E}[\|\Delta_0\|] + \eta K \|\nabla f_\eta(\boldsymbol{x}_0)\| \\
&\overset{(b)}{\leq} \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[\|\boldsymbol{x}_{i,K}^t - \boldsymbol{x}_t\|] + \eta K \|\nabla f_\eta(\boldsymbol{x}_0)\| \\
&\leq \frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \mathbb{E}[\|\boldsymbol{x}_{i,k+1}^t - \boldsymbol{x}_{i,k}^t\|] + \eta K \|\nabla f_\eta(\boldsymbol{x}_0)\| \\
&\leq \frac{\eta}{M}\sum_{i=1}^{M}\sum_{k=0}^{K} \mathbb{E}[\|g_{i,k}^t\|] + \eta K \|\nabla f_\eta(\boldsymbol{x}_0)\| \\
&\overset{(c)}{\leq} \eta K \Phi_\eta + \eta K \|\nabla f_\eta(\boldsymbol{x}_0)\| \\
&\overset{(d)}{\leq} \eta K \Phi_\eta + \eta K \|\nabla f(\boldsymbol{x}_0)\| + \frac{\eta^2 L (n+3)^{\frac{3}{2}}}{2} \\
&\overset{(e)}{\leq} \eta K \Phi_\eta + \eta K G + \frac{\eta^2 L (n+3)^{\frac{3}{2}}}{2} \\
&= \eta_0 \left(K\Phi_\eta + \frac{L\eta (n+3)^{\frac{3}{2}}}{2} + KG \right),
\end{aligned}$$

where (a) is due to the triangle inequality of norm, (b) follows from the definition of Δ, (c) is because of Lemma 3, (d) follows from Lemma 2, (e) is due to Assumption 3. Subsequently, we can give a formal bound for (45b):

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|] \leq T^{-\frac{1}{4}} \left(\frac{3\Delta_f}{\eta_0} + 8L\eta_0 + \sqrt{\frac{\Phi_\eta + \sigma^2}{K} + K^2 L^2 \Phi_\eta \eta_0^2} \right) + 8T^{-\frac{1}{2}} \left(G + \Phi_\eta + \frac{L\eta (n+3)^{\frac{3}{2}}}{2K} \right) \tag{46a}$$

$$+ T^{-\frac{3}{4}} \cdot (n+3)^{\frac{3}{2}} L\eta_0 + T^{-\frac{7}{4}} \cdot 3nL\eta_0 \tag{46b}$$

$$\leq T^{-\frac{1}{4}} \left(\frac{3\Delta_f}{\eta_0} + 8L\eta_0 + \sqrt{\frac{\Phi_\eta + \sigma^2}{K} + K^2 L^2 \Phi_\eta \eta_0^2} \right) + T^{-\frac{1}{2}} \cdot 8(G + \Phi_\eta) + T^{-\frac{3}{4}} \cdot \eta_0 L (n+3)^{\frac{3}{2}} \tag{46c}$$

$$+ T^{-\frac{5}{4}} \frac{4L\eta_0 (n+3)^{\frac{3}{2}}}{K} + T^{-\frac{7}{4}} \cdot 3nL\eta_0. \tag{46d}$$

Let $\Phi := \Phi_\eta$, where $\eta := \frac{\eta_0}{T^{\frac{3}{4}}}$. By Lemma 3,

$$\Phi = T^{-\frac{3}{2}} \frac{L^2 \eta_0^2}{2} (n+6)^3 + 10n(G^2 + \sigma^2).$$

■