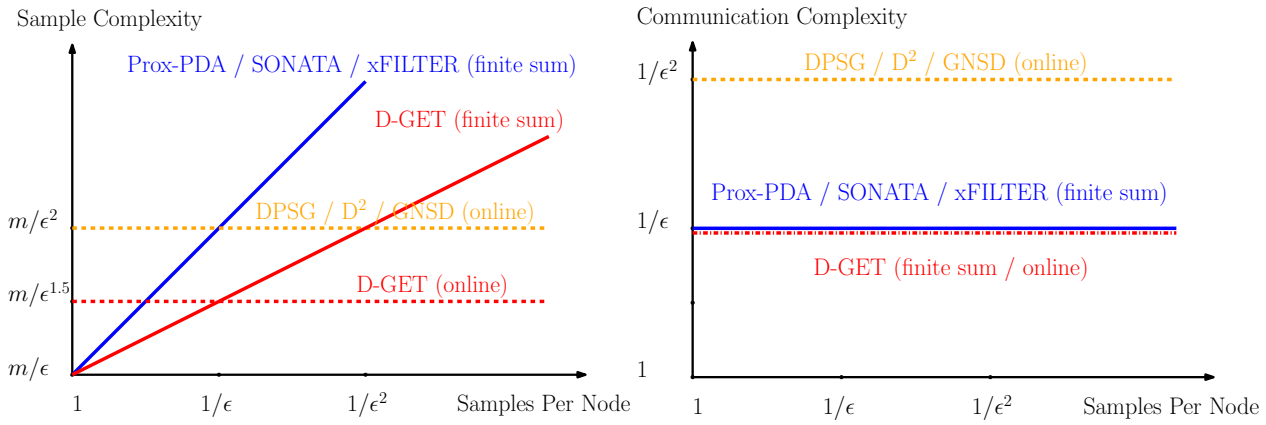



(a) Sample Complexity (b) Communication Complexity

*Figure 3.* Comparison of the sample and communication complexities for a number of decentralized methods. Existing deterministic methods enjoy lower sample complexity at smaller sample sizes, but such complexity scales linearly when the number of samples increases. Stochastic methods generally suffer from high communication complexity. The proposed D-GET bridges the gap between existing deterministic and stochastic methods, and achieves the optimal sample and communication complexities. Note that online methods can also be applied for finite sum problems, thus the actual sample complexity of D-GET is the minimum rate of both cases.

## A. Choices of Mixing Matrices

Note that many choices of mixing matrices satisfy the above condition. Here we give three commonly used mixing matrices (Xiao & Boyd, 2004; Boyd et al., 2004), where $d_i$ denotes the degree of node $i$, and $d_{\max} = \max_i\{d_i\}$:

- Metropolis-Hasting Weight

$$w_{ij} = \begin{cases} \frac{1}{1+\max\{d_i,d_j\}}, & \text{if} \quad \{i,j\} \in \mathcal{E}, \\ 1 - \sum_{\{i,k\}\in\mathcal{E}} w_{ik}, & \text{if} \quad i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

- Maximum-Degree Weight

$$w_{ij} = \begin{cases} \frac{1}{d_{\max}}, & \text{if} \quad \{i,j\} \in \mathcal{E}, \\ 1 - \frac{d_i}{d_{\max}}, & \text{if} \quad i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

- Laplacian Weight

$$w_{ij} = \begin{cases} \gamma & \text{if} \quad \{i,j\} \in \mathcal{E}, \\ 1 - \gamma d_i, , & \text{if} \quad i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

If we use $\mathcal{L}$ to denote the graph Laplacian matrix, and $\lambda_{\max}, \lambda_{\min}$ as the largest and second smallest eigenvalue, then one of the common choices of $\gamma$ is $\frac{2}{\lambda_{\max}(\mathcal{L})+\lambda_{\min}(\mathcal{L})}$.

## B. Convergence Analysis

In the appendix, we provide a complete theoretical analysis on the convergence of the proposed D-GET algorithm.

Before we formally conduct the analysis, we note three simple facts about Algorithm 1.

First, according to (9) and the definition (13a), the update rule of the average iterates can be expressed as:

$$\bar{\mathbf{x}}^r = \bar{\mathbf{x}}^{r-1} - \alpha \bar{\mathbf{y}}^{r-1}. \tag{32}$$

Second, if the iteration $r$ satisfies $\text{mod}(r, q) = 0$ (that is, when the outer iteration is executed), from (10) and (12) it is easy to check that the following relations hold (given $\mathbf{v}^0 = \mathbf{y}^0$):

$$\bar{\mathbf{v}}^r = \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r) = \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n \nabla f_k^i(\mathbf{x}_i^r), \tag{33}$$

$$\bar{\mathbf{y}}^r = \bar{\mathbf{v}}^r, \text{ if } \text{mod}(r, q) = 0. \tag{34}$$

Third, if $\text{mod}(r, q) \neq 0$, we have the following relations:

$$\bar{\mathbf{v}}^r = \frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\left[\nabla f_j^i(\mathbf{x}_i^r) - \nabla f_j^i(\mathbf{x}_i^{r-1})\right] + \bar{\mathbf{v}}^{r-1}, \tag{35}$$

$$\bar{\mathbf{y}}^r = \bar{\mathbf{y}}^{r-1} + \bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1}, \text{ if } \text{mod}(r, q) \neq 0. \tag{36}$$

### B.1. Proof of Lemma 1

*Proof.* Define $\mathbb{E}[\cdot|\mathcal{F}_r]$ as the expectation with respect to the random choice of sample $j$, conditioned on $\mathbf{x}^0, \cdots, \mathbf{x}^r$, $\mathbf{v}^0, \cdots, \mathbf{v}^{r-1}$ and $\mathbf{y}^0, \cdots, \mathbf{y}^{r-1}$.

First, we have the following identity (which holds true when $\text{mod}(r, q) \neq 0$)

$$\mathbb{E}[\bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1}|\mathcal{F}_r] \overset{(35)}{=} \mathbb{E}\left[\frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\left[\nabla f_j^i(\mathbf{x}_i^r) - \nabla f_j^i(\mathbf{x}_i^{r-1})\right]\Big|\mathcal{F}_r\right] = \frac{1}{m}\sum_{i=1}^m\left[\nabla f^i(\mathbf{x}_i^r) - \nabla f^i(\mathbf{x}_i^{r-1})\right]. \tag{37}$$

To see why the second equality holds, note that when $\mathbf{x}^r, \mathbf{y}^{r-1}, \mathbf{v}^{r-1}$ are known and fixed, the second expectation is taken over the random selection of $S_2$. The second equality follows because $S_2$ is sampled from $[n]$ uniformly with replacement, and it is an unbiased estimate of the averaged gradient.

Second, it is straightforward to obtain the following equality,

$$\begin{aligned}\left\|\bar{\mathbf{y}}^r - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\right\|^2 &\overset{(36)}{=} \left\|\bar{\mathbf{y}}^{r-1} + \bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\right\|^2 \\ &= \left\|\bar{\mathbf{y}}^{r-1} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1})\right\|^2 + \left\|\bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1} + \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1}) - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\right\|^2 \\ &\quad + 2\left\langle \bar{\mathbf{y}}^{r-1} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1}), \bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1} + \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1}) - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\right\rangle,\end{aligned} \tag{38}$$

where in the second equality, we add and subtract a term $\frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1})$.

The cross term in (38) can be eliminated if we take the conditional expectation conditioning on $\mathcal{F}_r$. Since under $\mathcal{F}_r$, we have $\mathbf{x}^r, \mathbf{x}^{r-1}, \mathbf{v}^{r-1}, \mathbf{y}^{r-1}, \bar{\mathbf{y}}^{r-1}$ are all known and fixed. Further applying (37) we have

$$\mathbb{E}\left[\bar{\mathbf{v}}^r - \bar{\mathbf{v}}^{r-1} + \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^{r-1}) - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\Big|\mathcal{F}_r\right] = 0. \tag{39}$$

Further taking the full expectation on (38) we have

$$\mathbb{E}\left\|\bar{\mathbf{y}}^r - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\right\|^2 \overset{(2)(35)(39)}{=} \mathbb{E}\left\|\bar{\mathbf{y}}^{r-1} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i{}^{r-1})\right\|^2 \tag{40}$$
$$+\mathbb{E}\left\|\frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\nabla f_j^i(\mathbf{x}_i^{r-1}) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r)\right\|^2.$$

Since each $j \in S_2$ is sampled from $[n]$ uniformly, we have the following conditional expectation:

$$\mathbb{E}_j\left[\frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1})\middle|\mathcal{F}_r\right] = 0. \tag{41}$$

Then the second term of RHS of (40) can be further bounded through following (where $\mathbb{E}[\cdot]$ is the full expectation)

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m|S_2|}\sum_{i=1}^m\sum_{j\in S_2}\nabla f_j^i(\mathbf{x}_i^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1})\right\|^2\\
=&\frac{1}{|S_2|^2}\mathbb{E}\left\|\sum_{j\in S_2}\left(\frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1})\right)\right\|^2\\
\overset{(i)}{=}&\frac{1}{|S_2|^2}\sum_{j\in S_2}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1})\right\|^2\\
\overset{(ii)}{\le}&\frac{1}{|S_2|}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1})\right\|^2\\
\overset{(iii)}{\le}&\frac{1}{m|S_2|}\mathbb{E}\left[\sum_{i=1}^m\|\nabla f_j^i(\mathbf{x}_i^r) - \nabla f_j^i(\mathbf{x}_i^{r-1})\|^2\right]\\
\overset{(iv)}{\le}&\frac{L^2}{m|S_2|}\mathbb{E}\|\mathbf{x}^r - \mathbf{x}^{r-1}\|^2,
\end{aligned} \tag{42}$$

In step $(i)$ of the above relation, we use the fact that for two random variables $u_i$, $u_j$ which are independent conditioning on $\mathcal{F}$, the following holds

$$\mathbb{E}[\langle u_\ell, u_j\rangle] = \mathbb{E}_{\mathcal{F}}\mathbb{E}[\langle u_\ell, u_j\rangle \mid \mathcal{F}] = \mathbb{E}_{\mathcal{F}}\langle\mathbb{E}[u_\ell \mid \mathcal{F}], \mathbb{E}[u_j \mid \mathcal{F}]\rangle. \tag{43}$$

Plugging $u_j$ as below and $u_\ell$ similarly,

$$u_j = \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1}) - \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i^r) + \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\nabla f_k^i(\mathbf{x}_i{}^{r-1}) \tag{44}$$

and note that (41) holds true, we can show that the cross terms in the step before $(i)$ can all be eliminated. In step $(ii)$ of (42), we use the property that $\mathbb{E}\|w_j - \mathbb{E}(w_j)\|^2 \le \mathbb{E}\|w_j\|^2$ and $\mathbb{E}\|w_j\|^2 = \mathbb{E}\|w_k\|^2$ with $w_j := \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^m\nabla f_j^i(\mathbf{x}_i^{r-1})$; in $(iii)$ we use the Jensen's inequality, and the last inequality $(iv)$ follows Assumption 1.

Therefore, by combining (40) and (42), we have for all $(n_r - 1)q + 1 \le r \le n_r q - 1$,

$$\mathbb{E}\left\|\bar{\mathbf{y}}^r - \frac{1}{m}\sum_{i=1}^m\nabla f^i(\mathbf{x}_i^r)\right\|^2 \le \mathbb{E}\left\|\bar{\mathbf{y}}^{r-1} - \frac{1}{m}\sum_{i=1}^m\nabla f^i(\mathbf{x}_i{}^{r-1})\right\|^2 + \frac{L^2}{m|S_2|}\mathbb{E}\left\|\mathbf{x}^r - \mathbf{x}^{r-1}\right\|^2. \tag{45}$$

Next, note that we have the following bound on $\mathbb{E}\|\mathbf{x}^r - \mathbf{x}^{r-1}\|^2$ for all $r \geq 1$:

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^r - \mathbf{x}^{r-1}\|^2 &\overset{(9)}{=} \mathbb{E}\|\mathbf{W}\mathbf{x}^{r-1} - \alpha\mathbf{y}^{r-1} - \mathbf{x}^{r-1}\|^2 \\ &\overset{(i)}{\leq} 2\mathbb{E}\|\mathbf{W}\mathbf{x}^{r-1} - \mathbf{x}^{r-1}\|^2 + 2\alpha^2\mathbb{E}\|\mathbf{y}^{r-1}\|^2 \\ &\overset{(ii)}{\leq} 2\mathbb{E}\|(\mathbf{W}-\mathbf{I})(\mathbf{x}^{r-1} - \mathbf{1}\bar{\mathbf{x}}^{r-1})\|^2 + 2\alpha^2\mathbb{E}\|\mathbf{y}^{r-1}\|^2 \\ &\overset{(iii)}{\leq} 8\mathbb{E}\|\mathbf{x}^{r-1} - \mathbf{1}\bar{\mathbf{x}}^{r-1}\|^2 + 4\alpha^2\mathbb{E}\|\mathbf{y}^{r-1} - \mathbf{1}\bar{\mathbf{y}}^{r-1}\|^2 + 4\alpha^2\mathbb{E}\|\mathbf{1}\bar{\mathbf{y}}^{r-1}\|^2, \ \forall\, r \geq 1 \end{aligned} \tag{46}$$

where in $(i)$ we apply the Cauchy-Schwarz inequality, $(ii)$ follows that $\mathbf{W}\mathbf{1} = \mathbf{1}$ from Assumption 2, and $(iii)$ applies the fact $\|\mathbf{W} - \mathbf{I}\| \leq \|\mathbf{W}\| + \|\mathbf{I}\| \leq 2$ (due to Assumption 2 and the Cauchy-Schwarz inequality).

Telescoping the above inequality (45) over the $n_r$-th inner loop, that is from $(n_r-1)q+1$ to $r$, we obtain the following series of inequalities

$$\begin{aligned}&\mathbb{E}\|\bar{\mathbf{y}}^r - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i^r)\|^2 \\ \leq&\frac{L^2}{m|S_2|}\sum_{t=(n_r-1)q+1}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2 + \mathbb{E}\|\bar{\mathbf{y}}^{(n_r-1)q} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i{}^{(n_r-1)q})\|^2 \\ \overset{(46)}{\leq}&\frac{8L^2}{m|S_2|}\sum_{t=(n_r-1)q+1}^{r}\mathbb{E}\|\mathbf{x}^{t-1} - \mathbf{1}\bar{\mathbf{x}}^{t-1}\|^2 + \frac{4\alpha^2L^2}{m|S_2|}\sum_{t=(n_r-1)q+1}^{r}\mathbb{E}\|\mathbf{y}^{t-1} - \mathbf{1}\bar{\mathbf{y}}^{t-1}\|^2 \\ &+ \frac{4\alpha^2L^2}{m|S_2|}\sum_{t=(n_r-1)q+1}^{r}\mathbb{E}\|\mathbf{1}\bar{\mathbf{y}}^{t-1}\|^2 + \mathbb{E}\|\bar{\mathbf{y}}^{(n_r-1)q} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i{}^{(n_r-1)q})\|^2 \\ \overset{(i)}{\leq}&\frac{8L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{4\alpha^2L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 \\ &+ \frac{4\alpha^2L^2}{|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \mathbb{E}\|\bar{\mathbf{y}}^{(n_r-1)q} - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\mathbf{x}_i{}^{(n_r-1)q})\|^2,\end{aligned}$$

where in $(i)$ we change the index in the summation, and add three non-negative terms (one for each sum). This concludes the first part of this lemma.

Next we show that (15) holds true. First, by using the same argument as in (37), we can obtain the following

$$\mathbb{E}[\mathbf{v}^r - \mathbf{v}^{r-1}|\mathcal{F}_r] \overset{(10)}{=} \mathbb{E}\left[\frac{1}{|S_2|}\sum_{j\in S_2}\left[\nabla f_j(\mathbf{x}^r) - \nabla f_j(\mathbf{x}^{r-1})\right]\Big|\mathcal{F}_r\right] = \nabla f(\mathbf{x}^r) - \nabla f(\mathbf{x}^{r-1}).$$

By using the above fact, and that conditioning on $\mathcal{F}_r$, $\mathbf{x}^r, \mathbf{x}^{r-1}$ and $\mathbf{v}^{r-1}$, we obtain the following:

$$\mathbb{E}[\langle \mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1}), \mathbf{v}^r - \mathbf{v}^{r-1} - \nabla f(\mathbf{x}^r) + \nabla f(\mathbf{x}^{r-1})\rangle|\mathcal{F}_r] = 0. \tag{47}$$

Then it is straightforward to obtain following:

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^r - \nabla f(\mathbf{x}^r)\|^2 &= \mathbb{E}\|\mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1}) + \mathbf{v}^r - \mathbf{v}^{r-1} - \nabla f(\mathbf{x}^r) + \nabla f(\mathbf{x}^{r-1})\|^2 \\ &\overset{(47)}{=} \mathbb{E}\|\mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1})\|^2 + \mathbb{E}\|\mathbf{v}^r - \mathbf{v}^{r-1} - \nabla f(\mathbf{x}^r) + \nabla f(\mathbf{x}^{r-1})\|^2 \\ &\overset{(10)}{=} \mathbb{E}\|\mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1})\|^2 + \mathbb{E}\|\frac{1}{|S_2|}\sum_{j\in S_2}\nabla f_j(\mathbf{x}^r) - \frac{1}{|S_2|}\sum_{j\in S_2}\nabla f_j(\mathbf{x}^{r-1}) - \nabla f(\mathbf{x}^r) + \nabla f(\mathbf{x}^{r-1})\|^2 \\ &\overset{(i)}{\leq} \mathbb{E}\|\mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1})\|^2 + \frac{1}{|S_2|}\mathbb{E}\|\nabla f_j(\mathbf{x}^r) - \nabla f_j(\mathbf{x}^{r-1})\|^2 \\ &\overset{(ii)}{\leq} \mathbb{E}\|\mathbf{v}^{r-1} - \nabla f(\mathbf{x}^{r-1})\|^2 + \frac{L^2}{|S_2|}\mathbb{E}\|\mathbf{x}^r - \mathbf{x}^{r-1}\|^2,\end{aligned}$$

where (i) and (ii) follow the similar arguments as in (42).

Telescoping the above inequality over $r$ from $(n_r-1)q+1$ to $r$, we obtain that

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^r-\nabla f(\mathbf{x}^r)\|^2 &\le \mathbb{E}\|\mathbf{v}^{(n_r-1)q}-\nabla f(\mathbf{x}^{(n_r-1)q})\|^2+\frac{L^2}{|S_2|}\sum_{t=(n_r-1)q+1}^{r}\mathbb{E}\|\mathbf{x}^t-\mathbf{x}^{t-1}\|^2\\ &\le \mathbb{E}\|\mathbf{v}^{(n_r-1)q}-\nabla f(\mathbf{x}^{(n_r-1)q})\|^2+\frac{L^2}{|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2.\end{aligned}$$

This completes the proof of the second part of this lemma. □

### B.2. Proof of Lemma 2

*Proof.* We first establish the relation of function values between the iterates. According to the gradient Lipschitz continuity (Assumption 1), we have

$$\begin{aligned}f(\bar{\mathbf{x}}^{r+1}) &\le f(\bar{\mathbf{x}}^r)+\langle\nabla f(\bar{\mathbf{x}}^r),\bar{\mathbf{x}}^{r+1}-\bar{\mathbf{x}}^r\rangle+\frac{L}{2}\|\bar{\mathbf{x}}^{r+1}-\bar{\mathbf{x}}^r\|^2\\ &\overset{(i)}{=} f(\bar{\mathbf{x}}^r)-\alpha\langle\nabla f(\bar{\mathbf{x}}^r),\bar{\mathbf{y}}^r\rangle+\frac{\alpha^2L}{2}\|\bar{\mathbf{y}}^r\|^2\\ &\overset{(ii)}{=} f(\bar{\mathbf{x}}^r)-\alpha\langle\nabla f(\bar{\mathbf{x}}^r)-\bar{\mathbf{y}}^r,\bar{\mathbf{y}}^r\rangle-\alpha\|\bar{\mathbf{y}}^r\|^2+\frac{\alpha^2L}{2}\|\bar{\mathbf{y}}^r\|^2\\ &\overset{(iii)}{\le} f(\bar{\mathbf{x}}^r)+\frac{\alpha}{2}\|\nabla f(\bar{\mathbf{x}}^r)-\bar{\mathbf{y}}^r\|^2-\frac{\alpha}{2}\|\bar{\mathbf{y}}^r\|^2+\frac{\alpha^2L}{2}\|\bar{\mathbf{y}}^r\|^2\\ &\overset{(iv)}{\le} f(\bar{\mathbf{x}}^r)-\left(\frac{\alpha}{2}-\frac{\alpha^2L}{2}\right)\|\bar{\mathbf{y}}^r\|^2+\alpha\|\nabla f(\bar{\mathbf{x}}^r)-\frac{1}{m}\sum_{i=1}^m\nabla f^i(\mathbf{x}_i^r)\|^2+\alpha\|\frac{1}{m}\sum_{i=1}^m\nabla f^i(\mathbf{x}_i^r)-\bar{\mathbf{y}}^r\|^2,\end{aligned}$$

where we simply plug in the iterates (32) in $(i)$, add and subtract a term $\bar{\mathbf{y}}^r$ in $(ii)$, and apply the Cauchy-Schwarz inequality in $(iii)$ and $(iv)$.

Then the third term can be further quantified as below,

$$\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)-\frac{1}{m}\sum_{i=1}^m\nabla f^i(\mathbf{x}_i^r)\|^2\overset{(i)}{\le}\frac{1}{m}\sum_{i=1}^m\mathbb{E}\|\nabla f^i(\bar{\mathbf{x}}^r)-\nabla f^i(\mathbf{x}_i^r)\|^2\overset{(ii)}{\le}\frac{1}{m}\sum_{i=1}^m L^2\mathbb{E}\|\mathbf{x}_i^r-\bar{\mathbf{x}}_i^r\|^2=\frac{L^2}{m}\mathbb{E}\|\mathbf{x}^r-\mathbf{1}\bar{\mathbf{x}}^r\|^2$$

where in $(i)$ we use the Jensen's inequality and in $(ii)$ we use the Lipschitz Assumption 1.

Taking expectation on both sides and combining with (14) in Lemma 1, we have

$$\begin{aligned}\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] &\le \mathbb{E}[f(\bar{\mathbf{x}}^r)]-\left(\frac{\alpha}{2}-\frac{\alpha^2L}{2}\right)\mathbb{E}\|\bar{\mathbf{y}}^r\|^2+\frac{\alpha L^2}{m}\mathbb{E}\|\mathbf{x}^r-\mathbf{1}\bar{\mathbf{x}}^r\|^2+\frac{8\alpha L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{x}^t-\mathbf{1}\bar{\mathbf{x}}^t\|^2\\ &\quad+\frac{4\alpha^3L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{y}^t-\mathbf{1}\bar{\mathbf{y}}^t\|^2+\frac{4\alpha^3L^2}{|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2+\alpha\mathbb{E}\|\bar{\mathbf{y}}^{(n_r-1)q}-\frac{1}{m}\sum_{t=1}^m\nabla f^t(\mathbf{x}_t{}^{(n_r-1)q})\|^2\\ &\le \mathbb{E}[f(\bar{\mathbf{x}}^r)]-\left(\frac{\alpha}{2}-\frac{\alpha^2L}{2}\right)\mathbb{E}\|\bar{\mathbf{y}}^r\|^2+\frac{\alpha L^2}{m}\mathbb{E}\|\mathbf{x}^r-\mathbf{1}\bar{\mathbf{x}}^r\|^2+\frac{8\alpha L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{x}^t-\mathbf{1}\bar{\mathbf{x}}^t\|^2\\ &\quad+\frac{4\alpha^3L^2}{m|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{y}^t-\mathbf{1}\bar{\mathbf{y}}^t\|^2+\frac{4\alpha^3L^2}{|S_2|}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2+\alpha\epsilon_1,\end{aligned}$$

where in the last inequality we use the definition of $\epsilon_1$ in (16).

Next, telescoping over one inner loop, that is $r$ from $(n_r-1)q$ to $r$, we have

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] \leq & \mathbb{E}[f(\bar{\mathbf{x}}^{(n_r-1)q})] - \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2}\right) \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \frac{\alpha L^2}{m} \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
& + \frac{8\alpha L^2}{m|S_2|} \sum_{t=(n_r-1)q}^{r} \sum_{k=(n_r-1)q}^{t} \mathbb{E}\|\mathbf{x}^k - \mathbf{1}\bar{\mathbf{x}}^k\|^2 + \frac{4\alpha^3 L^2}{m|S_2|} \sum_{t=(n_r-1)q}^{r} \sum_{k=(n_r-1)q}^{t} \mathbb{E}\|\mathbf{y}^k - \mathbf{1}\bar{\mathbf{y}}^k\|^2 \\
& + \frac{4\alpha^3 L^2}{|S_2|} \sum_{t=(n_r-1)q}^{r} \sum_{k=(n_r-1)q}^{t} \mathbb{E}\|\bar{\mathbf{y}}^k\|^2 + \alpha \sum_{t=(n_r-1)q}^{r} \epsilon_1 \\
\overset{(i)}{\leq} & \mathbb{E}[f(\bar{\mathbf{x}}^{(n_r-1)q})] - \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2} - \frac{4\alpha^3 L^2 q}{|S_2|}\right) \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \frac{\alpha L^2}{m} \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
& + \frac{8\alpha L^2 q}{m|S_2|} \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{4\alpha^3 L^2 q}{m|S_2|} \sum_{t=(n_r-1)q}^{r} \mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \alpha \sum_{t=(n_r-1)q}^{r} \epsilon_1,
\end{aligned}$$

where $(i)$ follows the fact that, for any sequence $\{a^i\}$, and an index $r \leq n_r q - 1$, we have

$$\sum_{t=(n_r-1)q}^{r} \sum_{k=(n_r-1)q}^{t} a^k \leq \sum_{t=(n_r-1)q}^{r} \sum_{k=(n_r-1)q}^{r} a^k \leq q \sum_{k=(n_r-1)q}^{r} a^k. \tag{48}$$

Then utilizing the fact that

$$\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^0)] = \mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^{(n_r-1)q})] + \cdots + \mathbb{E}[f(\bar{\mathbf{x}}^{2q})] - \mathbb{E}[f(\bar{\mathbf{x}}^q)] + \mathbb{E}[f(\bar{\mathbf{x}}^q)] - \mathbb{E}[f(\bar{\mathbf{x}}^0)], \tag{49}$$

we have

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] \leq & \mathbb{E}[f(\bar{\mathbf{x}}^0)] - \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2} - \frac{4\alpha^3 L^2 q}{|S_2|}\right) \sum_{t=0}^{r} \mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \frac{\alpha L^2}{m} \sum_{t=0}^{r} \mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
& + \frac{8\alpha L^2 q}{m|S_2|} \sum_{t=0}^{r} \mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{4\alpha^3 L^2 q}{m|S_2|} \sum_{t=0}^{r} \mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \alpha(r+1)\epsilon_1,
\end{aligned}$$

which completes the proof. □

### B.3. Proof of Lemma 3

*Proof.* First, using the Assumption 2 on $\mathbf{W}$, we can obtain the contraction property of the iterates disagreement, i.e.,

$$\|\mathbf{W}\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\| = \|\mathbf{W}(\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r)\| \leq \eta \|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|. \tag{50}$$

To see why the inequality holds true, note that $\mathbf{1}^T(\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r) = 0$, that is, $\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r$ is orthogonal $\mathbf{1}$, which is the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}$. Combining with the fact that $|\lambda_{\max}(\mathbf{W})| = \eta < 1$, we obtain the above inequality.

Then applying the definition of $\mathbf{x}$ iterates (9) and the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\|\mathbf{x}^{r+1} - \mathbf{1}\bar{\mathbf{x}}^{r+1}\|^2 \overset{(9)}{=} & \|\mathbf{W}\mathbf{x}^r - \alpha\mathbf{y}^r - \mathbf{1}(\bar{\mathbf{x}}^r - \alpha\bar{\mathbf{y}}^r)\|^2 \\
\leq & (1+\beta)\|\mathbf{W}\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 + \left(1 + \frac{1}{\beta}\right)\alpha^2 \|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 \\
\overset{(50)}{\leq} & (1+\beta)\eta^2 \|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 + \left(1 + \frac{1}{\beta}\right)\alpha^2 \|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2,
\end{aligned}$$

where $\beta$ is some constant parameter to be tuned later. Then, taking expectation on the both sides of the above inequality we are able to obtain (18).

Similarly, we have

$$\begin{aligned}\|\mathbf{y}^{r+1}-\mathbf{1}\bar{\mathbf{y}}^{r+1}\|^2 &\overset{(12)}{=} \|\mathbf{W}\mathbf{y}^r+\mathbf{v}^{r+1}-\mathbf{v}^r-\mathbf{1}(\bar{\mathbf{y}}^r+\bar{\mathbf{v}}^{r+1}-\bar{\mathbf{v}}^r)\|^2\\ &\le (1+\beta)\|\mathbf{W}\mathbf{y}^r-\mathbf{1}\bar{\mathbf{y}}^r\|^2+\left(1+\frac{1}{\beta}\right)\left\|\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^T}{m}\right)(\mathbf{v}^{r+1}-\mathbf{v}^r)\right\|^2\\ &\le (1+\beta)\eta^2\|\mathbf{y}^r-\mathbf{1}\bar{\mathbf{y}}^r\|^2+\left(1+\frac{1}{\beta}\right)\|\mathbf{v}^{r+1}-\mathbf{v}^r\|^2,\end{aligned}$$

where in the last inequality we also use $\|\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^T\|<1$.

After taking expectation on the both sides of the above inequality and combining the following inequalities, the proof for (19) is complete.

To further bound the term $\|\mathbf{v}^{r+1}-\mathbf{v}^r\|^2$, consider that we have $(n_r-1)q\le r\le n_rq-1$, that is $r$ is taken within one inner loop. We will divide the analysis into two cases.

**Case 1)** For all $(n_r-1)q\le r\le n_rq-2$, we have $\text{mod}(r+1,q)\neq 0$ and the following is straightforward:

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^{r+1}-\mathbf{v}^r\|^2 &\overset{(10)}{=} \mathbb{E}\|\frac{1}{|S_2|}\sum_{j\in S_2}\left[\nabla f_j(\mathbf{x}^{r+1})-\nabla f_j(\mathbf{x}^r)\right]\|^2\\ &\overset{(i)}{\le} \frac{1}{|S_2|}\mathbb{E}\sum_{j\in S_2}\|\nabla f_j(\mathbf{x}^{r+1})-\nabla f_j(\mathbf{x}^r)\|^2\\ &\overset{(ii)}{\le} L^2\mathbb{E}\|\mathbf{x}^{r+1}-\mathbf{x}^r\|^2,\end{aligned}\tag{51}$$

where in $(i)$ we use the Jensen's inequality and in $(ii)$ we use Assumption 1.

**Case 2)** If $r=n_rq-1$, we have $\text{mod}(r+1,q)=0$. Therefore,

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^{r+1}-\mathbf{v}^r\|^2 &= \mathbb{E}\|\mathbf{v}^{r+1}-\nabla f(\mathbf{x}^{r+1})+\nabla f(\mathbf{x}^{r+1})-\nabla f(\mathbf{x}^r)+\nabla f(\mathbf{x}^r)-\mathbf{v}^r\|^2\\ &\overset{(i)}{\le} 3\mathbb{E}\|\mathbf{v}^{r+1}-\nabla f(\mathbf{x}^{r+1})\|^2+3\mathbb{E}\|\nabla f(\mathbf{x}^{r+1})-\nabla f(\mathbf{x}^r)\|^2+3\mathbb{E}\|\nabla f(\mathbf{x}^r)-\mathbf{v}^r\|^2\\ &\overset{(ii)}{\le} 3\epsilon_2+3L^2\mathbb{E}\|\mathbf{x}^{r+1}-\mathbf{x}^r\|^2+3\sum_{t=(n_r-1)q}^{r}\frac{L^2}{|S_2|}\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+3\epsilon_2,\end{aligned}\tag{52}$$

where in $(i)$ we use the Cauchy-Schwarz inequality; in $(ii)$ we apply (15) from Lemma 1, Assumption 1, and $\mathbb{E}\|\mathbf{v}^r-\nabla f(\mathbf{x}^r)\|^2\le\epsilon_2$ for all $\text{mod}(r,q)=0$.

Next, telescoping $\|\mathbf{v}^{r+1}-\mathbf{v}^r\|^2$ over $r$ from $(n_r-1)q$ to $r$. Since $r\le n_rq-1$, we have at most one follows (52) and all the rest follow (51). Therefore, we obtain

$$\begin{aligned}\sum_{t=(n_r-1)q}^{r}\mathbb{E}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2 &\le \sum_{t=(n_r-1)q}^{r}L^2\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+6\epsilon_2+2L^2\mathbb{E}\|\mathbf{x}^{r+1}-\mathbf{x}^r\|^2+\sum_{t=(n_r-1)q}^{r}\frac{3L^2}{|S_2|}\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\\ &\le \sum_{t=(n_r-1)q}^{r}6L^2\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+6\epsilon_2.\end{aligned}$$

Through a similar step as (49), the following is obvious

$$\sum_{t=0}^{r}\mathbb{E}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2\le 6(r+1)\epsilon_2+\sum_{t=0}^{r}6L^2\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2.$$

By combining (46), i.e.,

$$\mathbb{E}\|\mathbf{x}^{r+1}-\mathbf{x}^r\|^2\le 8\mathbb{E}\|\mathbf{x}^r-\mathbf{1}\bar{\mathbf{x}}^r\|^2+4\alpha^2\mathbb{E}\|\mathbf{y}^r-\mathbf{1}\bar{\mathbf{y}}^r\|^2+4\alpha^2\mathbb{E}\|\mathbf{1}\bar{\mathbf{y}}^r\|^2,\ \forall\, r\ge 0,$$

we complete the proof. □

### B.4. Proof of Lemma 4

*Proof.* We first introduce an intermediate function $P(\mathbf{x}^r)$ to facilitate the analysis,

$$P(\mathbf{x}^r) := \mathbb{E}\|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 + \alpha\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2.$$

Obviously, we have $H(\mathbf{x}^r) = \mathbb{E}[f(\bar{\mathbf{x}}^r)] + \frac{1}{m}P(\mathbf{x}^r)$.

By applying (18) and (19) in Lemma 3 we have

$$\begin{aligned}
P(\mathbf{x}^{r+1}) - P(\mathbf{x}^r) \leq\ & (1+\beta)\eta^2\mathbb{E}\|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 + (1+\frac{1}{\beta})\alpha^2\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 + \alpha(1+\beta)\eta^2\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 \\
& + \alpha(1+\frac{1}{\beta})\mathbb{E}\|\mathbf{v}^{r+1} - \mathbf{v}^r\|^2 - \mathbb{E}\|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 - \alpha\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 \\
=\ & -\left(1 - (1+\beta)\eta^2\right)\mathbb{E}\|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 - \left(\alpha - \alpha(1+\beta)\eta^2 - (1+\frac{1}{\beta})\alpha^2\right)\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 \\
& + \alpha(1+\frac{1}{\beta})\mathbb{E}\|\mathbf{v}^{r+1} - \mathbf{v}^r\|^2.
\end{aligned}$$

Next, summing over the iteration from 0 to $r$ we obtain

$$\begin{aligned}
P(\mathbf{x}^{r+1}) - P(\mathbf{x}^0) \leq\ & -\left(1 - (1+\beta)\eta^2\right)\sum_{t=0}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
& - \left(\alpha - \alpha(1+\beta)\eta^2 - (1+\frac{1}{\beta})\alpha^2\right)\sum_{t=0}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 \\
& + \alpha(1+\frac{1}{\beta})\sum_{t=0}^{r}\mathbb{E}\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2.
\end{aligned} \tag{53}$$

If we further pick $q = |S_2|$, then Lemma 2 becomes

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})] \leq & \mathbb{E}[f(\bar{\mathbf{x}}^0)] - \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2} - 4\alpha^3L^2\right)\sum_{t=0}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 \\
& + \frac{9\alpha L^2}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{4\alpha^3L^2}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \alpha(r+1)\epsilon_1.
\end{aligned} \tag{54}$$

Similarly, equation (21) of Lemma 3 becomes

$$\sum_{t=0}^{r}\mathbb{E}\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \leq 48L^2\sum_{t=0}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + 24L^2\alpha^2\sum_{t=0}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + 24L^2\alpha^2\sum_{t=0}^{r}\mathbb{E}\|\mathbf{1}\bar{\mathbf{y}}^t\|^2 + 6(r+1)\epsilon_2. \tag{55}$$

Therefore, combine (53), (54) and (55), we have

$$\begin{aligned}
H(\mathbf{x}^{r+1}) - H(\mathbf{x}^0) \leq & -\left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2} - 4\alpha^3L^2 - 24(1+\frac{1}{\beta})\alpha^3L^2\right)\sum_{t=0}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 \\
& - \left(1 - (1+\beta)\eta^2 - 48\alpha(1+\frac{1}{\beta})L^2 - 9\alpha L^2\right)\frac{1}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
& - \left(\alpha - \alpha(1+\beta)\eta^2 - (1+\frac{1}{\beta})\alpha^2 - 24(1+\frac{1}{\beta})\alpha^3L^2 - 4\alpha^3L^2\right)\frac{1}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \alpha(r+1)(\epsilon_1 + 6\frac{1}{m}(1+\frac{1}{\beta})\epsilon_2).
\end{aligned}$$

This completes the proof. □

### B.5. Proof of Theorem 1

*Proof.* To begin with, we notice that by applying the update rule from Algorithm 1, then for all $\text{mod}(r, q) = 0$, the following holds true

$$\mathbb{E}\|\mathbf{v}^r - \nabla f(\mathbf{x}^r)\|^2 \overset{(10)}{=} 0, \tag{56}$$

$$\mathbb{E}\|\bar{\mathbf{y}}^r - \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\nabla f_k^i(\mathbf{x}_i^r)\|^2 \overset{(33)}{=} 0, \tag{57}$$

which implies $\epsilon_1 = \epsilon_2 = 0$ for Lemma 2, Lemma 3, and Lemma 4.

Next, if we further pick $\beta$ such that $1-(1+\beta)\eta^2 > 0$ and choose $0 < \alpha < \min\{K_1, K_2, K_3\}$, we can rewrite Lemma 4 as below

$$H(\mathbf{x}^{r+1}) - H(\mathbf{x}^0) \leq -C_1\sum_{t=0}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 - C_2\sum_{t=0}^{r}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 - C_3\sum_{t=0}^{r}\frac{1}{m}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2, \tag{58}$$

where $C_1 > 0, C_2 > 0, C_3 > 0$.

Therefore the upper bound of the optimality gap can be quantified as the following

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
\leq&\frac{2}{T}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \frac{2}{T}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{y}}^t - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2,
\end{aligned}$$

where we use the Cauchy-Schwarz inequality.

Applying (14) from Lemma 1 with $\mathbb{E}\|\bar{\mathbf{y}}^{(n_r-1)q} - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i{}^{(n_r-1)q})\|^2 = 0$, telescoping over $r$ from 0 to $T$ (follows similar reasoning as (48) and (49)), and using the choice of $|S_2| = q = \sqrt{n}$, we have

$$\sum_{t=0}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 \leq \frac{8L^2}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{4\alpha^2L^2}{m}\sum_{t=0}^{r}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + 4\alpha^2L^2\sum_{t=0}^{r}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2.$$

Combining the above two inequalities we can obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
\leq&\left(\frac{16L^2}{mT} + \frac{1}{mT}\right)\sum_{t=0}^{T}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{8\alpha^2L^2}{mT}\sum_{t=0}^{T}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \left(\frac{8\alpha^2L^2}{T} + \frac{2}{T}\right)\sum_{t=0}^{T}\|\bar{\mathbf{y}}^t\|^2.
\end{aligned}$$

Further combining with (58), we have

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \leq C_0\cdot\frac{H(\mathbf{x}^0) - H(\mathbf{x}^{T+1})}{T} \leq C_0\cdot\frac{\mathbb{E}[f(\mathbf{x}^0)] - \underline{f}}{T}, \tag{59}$$

where

$$C_0 := \left(\frac{8\alpha^2L^2+2}{C_1} + \frac{16L^2+1}{mC_2} + \frac{8\alpha^2L^2}{mC_3}\right),$$

and the last inequality follows from

$$\begin{aligned}
H(\mathbf{x}^0) &:= \mathbb{E}[f(\bar{\mathbf{x}}^0)] + \mathbb{E}\|\mathbf{x}^0 - \mathbf{1}\bar{\mathbf{x}}^0\|^2 + \alpha\mathbb{E}\|\mathbf{y}^0 - \mathbf{1}\bar{\mathbf{y}}^0\|^2 = \mathbb{E}[f(\bar{\mathbf{x}}^0)], \\
H(\mathbf{x}^r) &:= \mathbb{E}[f(\bar{\mathbf{x}}^r)] + \mathbb{E}\|\mathbf{x}^r - \mathbf{1}\bar{\mathbf{x}}^r\|^2 + \alpha\mathbb{E}\|\mathbf{y}^r - \mathbf{1}\bar{\mathbf{y}}^r\|^2 \geq \mathbb{E}[f(\bar{\mathbf{x}}^r)] \geq \underline{f}.
\end{aligned}$$

This completes the proof. □

### B.6. Proof of Corollary 1

*Proof.* If we pick $T = \lfloor C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0)-\underline{f}}{\epsilon} \rfloor + 1 \geq C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0)-\underline{f}}{\epsilon}$, then we can obtain following from Theorem 1

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2\\ &\leq C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0)-\underline{f}}{T} \leq \epsilon.\end{aligned}$$

Therefore, the total samples needed will be the sum of outer loop complexity ($\lceil \frac{T}{q} \rceil$ times full ($n$) gradient evaluations per node) plus inner loop complexity ($T$ times $|S_2|$ gradient evaluations per node), by letting $q = |S_2| = \sqrt{n}$, we conclude that the total samples needed are

$$\begin{aligned}&m \times \left( \lceil \frac{T}{q} \rceil \cdot n + T \cdot |S_2| \right)\\ &\leq m \times \left( \left( \frac{T}{\sqrt{n}} + 1 \right) n + T\sqrt{n} \right)\\ &\leq m \left( n + 2C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0)-\underline{f}}{\epsilon}\sqrt{n} + 2\sqrt{n} \right)\\ &= \mathcal{O}\left( m \times \left( n + \frac{\sqrt{n}}{\epsilon} \right) \right).\end{aligned}$$

This completes the proof. □

### B.7. Proof of Corollary 2

*Proof.* In previous proof of Theorem 1 and Corollary 1 we already show the convergence respect to *both* gradient size and consensus error

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2.$$

To show our results also holds true when the gradient metric is evaluated at the average of the output $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$, we only need following relations (by using Jensens inequality).

$$\begin{aligned}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\bar{\mathbf{x}})\|^2 =& \|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\bar{\mathbf{x}}) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i) + \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i)\|^2 &(60)\\ \overset{\text{(i)}}{\leq}& 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\bar{\mathbf{x}}) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i)\|^2 + 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i)\|^2 &(61)\\ \overset{\text{(ii)}}{\leq}& \frac{2}{m}\sum_{i=1}^{m}\|\nabla f^i(\bar{\mathbf{x}}) - \nabla f^i(\mathbf{x}_i)\|^2 + 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i)\|^2 &(62)\\ \overset{\text{(iii)}}{\leq}& \frac{2L^2}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i)\|^2 &(63)\end{aligned}$$

where in (i) we use the Cauchy-Swahitz inequality, in (ii) we use the Jensen's inequality to move the average outside the Euclidean norm, and the last inequality uses the Assumption 1. The proof is complete by further combining (59) in Theorem 1 and similar reasoning in Corollary 1. □

### B.8. Proof of Lemma 5

*Proof.* First recall the definition of $\mathbb{E}[\cdot|\mathcal{F}_r]$ in Lemma 1, which is the expectation with respect to the random choice of sample $\xi$, conditioning on $\mathbf{x}^0, \cdots, \mathbf{x}^r$, $\mathbf{v}^0, \cdots, \mathbf{v}^{r-1}$ and $\mathbf{y}^0, \cdots, \mathbf{y}^{r-1}$.

Let us define a random variable $u_\xi$ as below and $u_\ell$ similarly,

$$u_\xi = \frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r). \tag{64}$$

Note that $u_\xi$ and $u_\ell$ are independent random variables conditioning on $\mathcal{F}$. Further, we have the following from Assumption 3

$$\mathbb{E}_\xi\left[\frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\middle|\mathcal{F}_r\right] = 0. \tag{65}$$

Therefore we have

$$\mathbb{E}[\langle u_\xi, u_\ell\rangle] = \mathbb{E}_{\mathcal{F}}\mathbb{E}[\langle u_\xi, u_\ell\rangle \mid \mathcal{F}] = \mathbb{E}_{\mathcal{F}}\langle\mathbb{E}[u_\xi \mid \mathcal{F}], \mathbb{E}[u_\ell \mid \mathcal{F}]\rangle = 0. \tag{66}$$

Following the update rule from Algorithm 2, we have the following relations for all $\text{mod}(r, q) = 0$

$$\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{y}}^r - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\right\|^2 &\overset{(28)}{=} \mathbb{E}\left\|\frac{1}{m|S_1|}\sum_{i=1}^{m}\sum_{\xi\in S_1}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\right\|^2 \\
&\overset{(i)}{=} \frac{1}{|S_1|^2}\mathbb{E}\left\|\sum_{\xi\in S_1}\left(\frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\right)\right\|^2 \\
&\overset{(ii)}{=} \frac{1}{|S_1|^2}\mathbb{E}\sum_{\xi\in S_1}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\right\|^2 \\
&\overset{(iii)}{=} \frac{1}{|S_1|}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\right\|^2 \\
&\overset{(iv)}{\leq} \frac{1}{m|S_1|}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f_\xi^i(\mathbf{x}_i^r) - \nabla f^i(\mathbf{x}_i^r)\right\|^2 \\
&\leq \frac{\sigma^2}{|S_1|},
\end{aligned}$$

where in $(i)$ we take out the constant $|S_1|$; in $(ii)$ we eliminate the cross terms via (66); in $(iii)$ we use the fact that the following term

$$\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_\xi^i(\mathbf{x}_i^r) - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^r)\|^2 \tag{67}$$

are equal across different samples $\xi$; in $(iv)$ we use the Jensen's inequality, and the last inequality follows the Assumption 4.

Similarly, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}^r - \nabla f(\mathbf{x}^r)\|^2 &\overset{(27)}{=} \mathbb{E}\left\|\frac{1}{|S_1|}\sum_{\xi\in S_1}\nabla f_\xi(\mathbf{x}^r) - \nabla f(\mathbf{x}^r)\right\|^2 \\
&= \frac{1}{|S_1|}\mathbb{E}\|\nabla f_\xi(\mathbf{x}^r) - \nabla f(\mathbf{x}^r)\|^2 \\
&\leq \frac{1}{|S_1|}\sum_{i=1}^{m}\mathbb{E}\|\nabla f_\xi^i(\mathbf{x}^r) - \nabla f^i(\mathbf{x}^r)\|^2 \\
&\leq \frac{m\sigma^2}{|S_1|}.
\end{aligned}$$

This completes the proof. □

### B.9. Proof of Theorem 2

*Proof.* Note that it is easy to check that Lemma 1, Lemma 2, Lemma 3 and Lemma 4 still hold true. And the quantity $\epsilon_1$ and $\epsilon_2$ can be determined by Lemma 5, i.e., $\epsilon_1 = \frac{\sigma^2}{|S_1|}$ and $\epsilon_2 = \frac{m\sigma^2}{|S_1|}$. Therefore, Lemma 4 can be rewritten as below if we follow Algorithm 2,

$$H(\mathbf{x}^{r+1}) - H(\mathbf{x}^0) \leq -C_1 \sum_{t=0}^{r} \mathbb{E}\|\bar{\mathbf{y}}^t\|^2 - C_2 \sum_{t=0}^{r} \frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 - C_3 \sum_{t=0}^{r} \frac{1}{m}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \epsilon_3, \tag{68}$$

with $\epsilon_3 = \alpha(r+1)(1 + 6(1+\frac{1}{\beta}))\frac{\sigma^2}{|S_1|}$.

Therefore the upper bound of the optimality gap can be derived in a similar way as Theorem 1,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
\leq &\frac{2}{T}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{y}}^t\|^2 + \frac{2}{T}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{y}}^t - \frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
\leq &\left(\frac{16L^2}{mT} + \frac{1}{mT}\right)\sum_{t=0}^{T}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \frac{8\alpha^2L^2}{mT}\sum_{t=0}^{T}\mathbb{E}\|\mathbf{y}^t - \mathbf{1}\bar{\mathbf{y}}^t\|^2 + \left(\frac{8\alpha^2L^2}{T} + \frac{2}{T}\right)\sum_{t=0}^{T}\|\bar{\mathbf{y}}^t\|^2 + \frac{2}{T}\sum_{t=0}^{T}\frac{\sigma^2}{|S_1|}.
\end{aligned}$$

Further combining (68) we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \\
\leq &C_0\left(\frac{H(\mathbf{x}^0) - H(\mathbf{x}^{T+1}) + \epsilon_3}{T}\right) + \frac{2T+2}{T}\frac{\sigma^2}{|S_1|} \\
\leq &C_0 \cdot \frac{\mathbb{E}[f(\mathbf{x}^0)] - \underline{f}}{T} + C_0 \cdot \frac{\alpha(T+1)(7+\frac{6}{\beta})\sigma^2}{T|S_1|} + \frac{2T+2}{T}\frac{\sigma^2}{|S_1|}.
\end{aligned}$$

After picking $|S_1| = \frac{4C_0\alpha(7+\frac{6}{\beta})\sigma^2 + 8\sigma^2}{\epsilon}$, we complete the proof. □

### B.10. Proof of Corollary 2

*Proof.* If we pick the following constants for Algorithm 2:

$$\begin{aligned}
&|S_1| = \frac{4C_0\alpha(7+\frac{6}{\beta})\sigma^2 + 8\sigma^2}{\epsilon},\ q = |S_2| = \sqrt{|S_1|}, \\
&T = 2\lfloor C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0) - \underline{f}}{\epsilon}\rfloor + 2 \geq 2C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0) - \underline{f}}{\epsilon}
\end{aligned}$$

then from Theorem 2 we can obtain

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\nabla f^i(\mathbf{x}_i^t)\|^2 + \frac{1}{T}\sum_{t=0}^{T}\frac{1}{m}\mathbb{E}\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 \leq \underbrace{C_0 \cdot \frac{\mathbb{E}f(\mathbf{x}^0) - \underline{f}}{T}}_{\leq \frac{\epsilon}{2}} + \frac{\epsilon}{2} \leq \epsilon.$$

Therefore we have that the per-node sample evaluations are given as

$$\lceil\frac{T}{q}\rceil \cdot |S_1| + T \cdot |S_2| \leq \left(\frac{T}{\sqrt{|S_1|}} + 1\right)|S_1| + T\sqrt{|S_1|} = O(\frac{1}{\epsilon} + \frac{1}{\epsilon^{3/2}}).$$

This completes the proof. □

## C. Experiment Details

In this section, we provide more implementation details and results to our experiments.

### C.1. Objective Functions

First, two types of classical smooth non-convex problems are used in our simulation for evaluating the performance of the compared algorithms: a) decentralized logistic regression with a non-convex regularizer and b) non-convex robust linear regression, the detailed objective functions are stated as the following:

**a) Decentralized logistic regression with non-convex regularizer**

The problem is stated as follows

$$\min_{\mathbf{x}\in\mathbb{R}^{nd}} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\ell(\mathbf{x}_i^\top \mathbf{v}_j^i, \mathbf{y}_j^i) + \sum_{k=1}^{d}\frac{\alpha \mathbf{x}_{i,k}^2}{1+\mathbf{x}_{i,k}^2}\right], \tag{69}$$

where $\mathbf{v}_j^i \in \mathbb{R}^d, \forall i, j$ denote the features of node $i$ and sample $j$, $\mathbf{y}_j^i \in \{+1, -1\}, \forall i$ denote the labels of node $i$ and sample $j$, and the loss function is defined as the cross-entropy loss

$$\ell(\mathbf{x}_i^\top \mathbf{v}_j^i, \mathbf{y}_j^i) := -\mathbf{y}_j^i \log\left(\frac{1}{1+e^{-\mathbf{x}_i^\top \mathbf{v}_j^i}}\right) - \left(1-\mathbf{y}_j^i\right)\log\left(1-\frac{1}{1+e^{-\mathbf{x}_i^\top \mathbf{v}_j^i}}\right). \tag{70}$$

The regularization parameter for non-convex term is set to $\alpha = 0.01$ in our simulation.

**b) Non-convex robust linear regression**

The problem is stated as follows

$$\min_{\mathbf{x}\in\mathbb{R}^{nd}} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\ell(\mathbf{y}_j^i - \mathbf{x}_i^T \mathbf{v}_j^i), \tag{71}$$

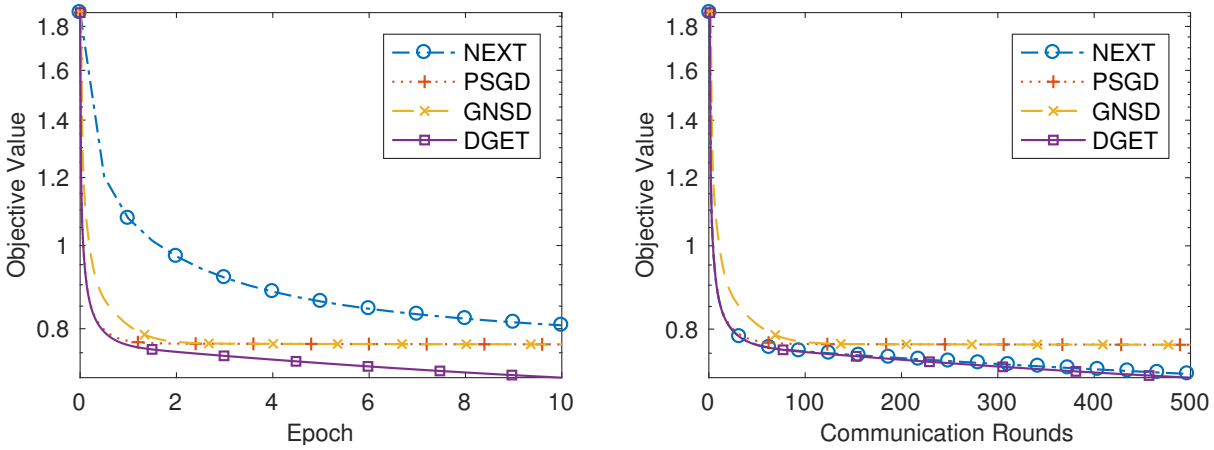
where $\mathbf{v}_j^i \in \mathbb{R}^d$ denotes the features of node $i$ and sample $j$, $\mathbf{y}_i \in \{+1, -1\}$ denotes the labels of node $i$ and sample $j$, and the loss function is defined as the following

$$\ell(u) := \log\left(\frac{u^2}{2}+1\right). \tag{72}$$

### C.2. Additional Simulations

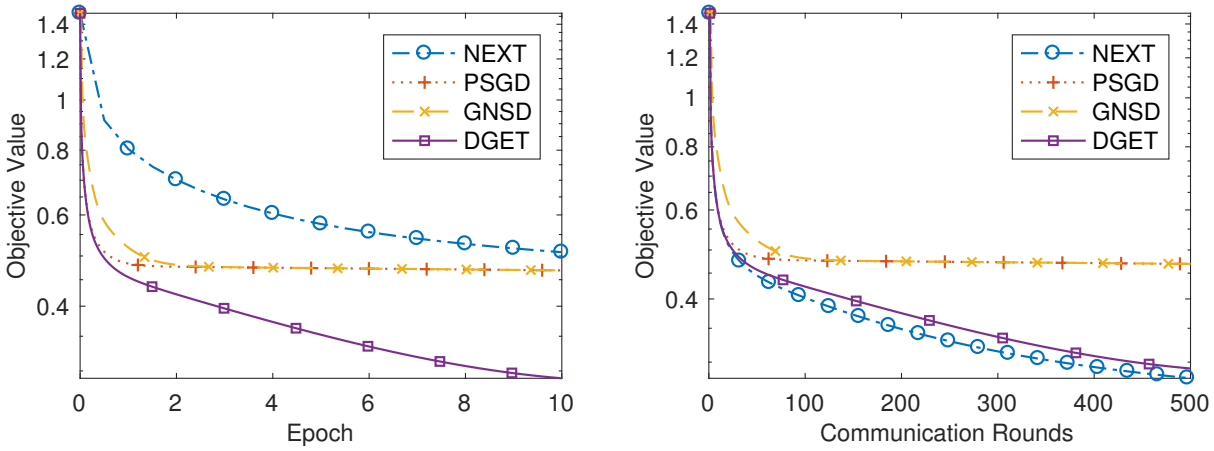
Next, we provide additional experimental results on the above mentioned decentralized logistic regression and robust linear regression problems. In particular, we demonstrate the performance of the algorithms in terms of the loss function as stated in (69) and (71) and the optimality gap defined in (4). For all algorithms considered, we set their learning rates to be $0.01$. For each experiment, we initialize all the algorithms at the same point generated randomly from the normal distribution. Also, we choose a fixed mini-batch size $64$ and set the epoch length $q$ to be $n/64$ such that all algorithms pass over the entire dataset once in each epoch.

The simulation results in terms of both sample complexity and the communication complexity averaged over 10 realizations on the a9a dataset $(n = 32561, d = 123)$ are shown in Figure 4 and Figure 5, and the performance on the w8a dataset $(n = 49749, d = 300)$ are reported in Figure 6 and Figure 7. It can be observed that the proposed D-GET could achieve much faster convergence in terms of sample complexity, while matches the communication complexity as the deterministic algorithms, as claimed in Theorem 1 and 2.



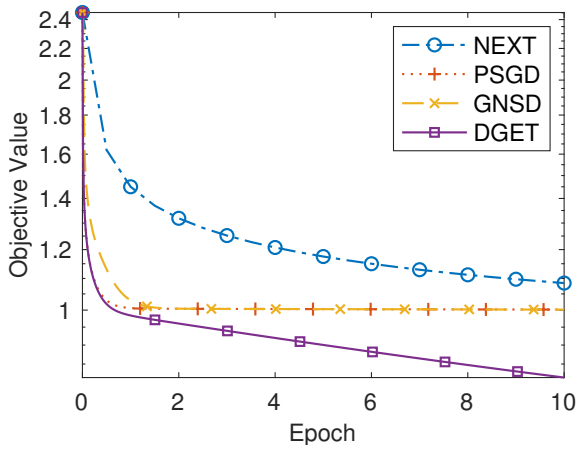

(a) Loss Function vs Sample Complexity (b) Loss Function vs Communication Complexity

*Figure 4.* Logistic regression on the a9a dataset over the path graph

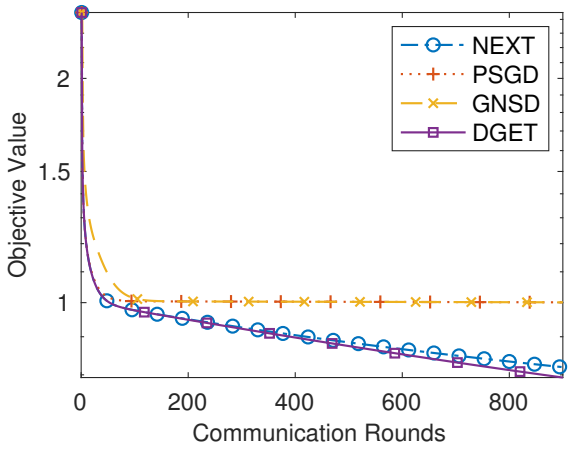



(a) Loss Function vs Sample Complexity (b) Loss Function vs Communication Complexity

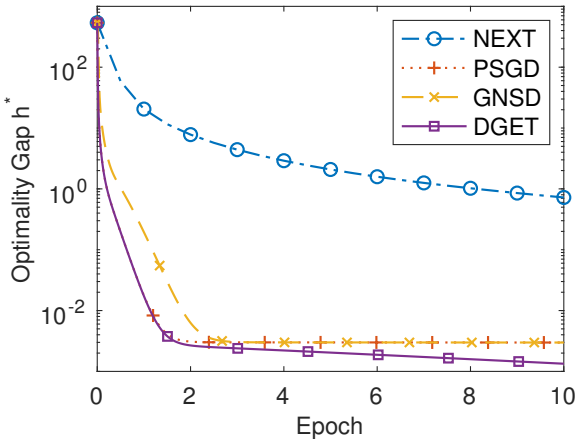
*Figure 5.* Robust linear regression on the a9a datatset over the path graph



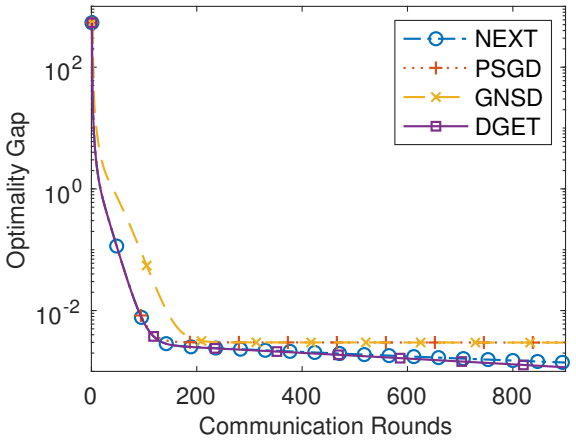

(a) Loss Function vs Sample Complexity




(b) Loss Function vs Communication Complexity

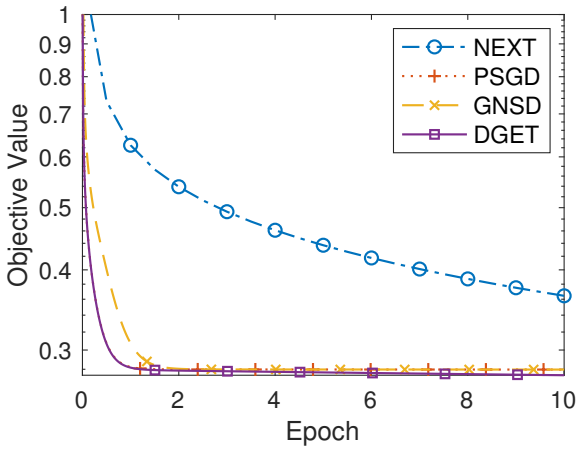



(c) Optimality Gap vs Sample Complexity

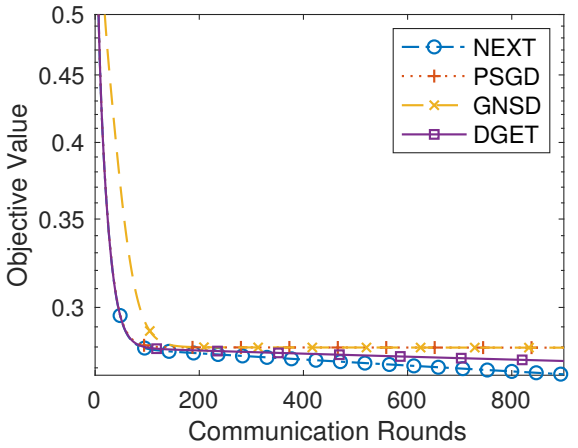



(d) Optimality Gap vs Communication Complexity

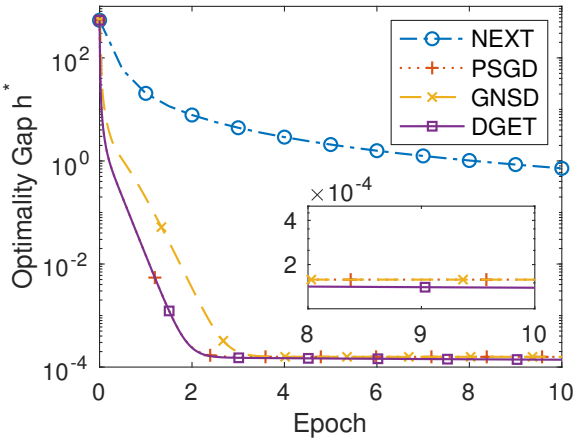
*Figure 6.* Logistic regression on the w8a dataset over the path graph



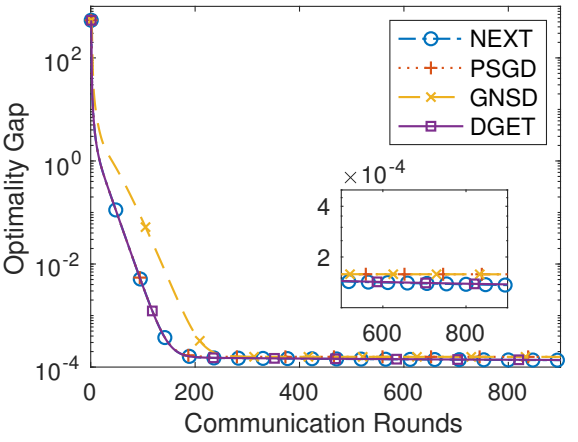

(a) Loss Function vs Sample Complexity




(b) Loss Function vs Communication Complexity




(c) Optimality Gap vs Sample Complexity




(d) Optimality Gap vs Communication Complexity

*Figure 7.* Robust linear regression on the w8a dataset over the path graph