APPENDIX

8 PROOFS

8.1 Useful Facts

We first derive some useful facts that will be used in the following proofs.

We can re-write Eq. (4) to

$$X_t = X_0 \bigsqcup_{s=0}^{t}(\neg M_s + M_s W_s) - \sum_{s=0}^{t-1}\gamma_s G(X_s;\xi_s) \bigsqcup_{r=s+1}^{t-1}(\neg M_r + M_r W_r). \quad (10)$$

Lemma 1. For any row vector sequence $\mathbf{x}_t \in \mathbb{R}^n$ defined as

$$\mathbf{x}_t = \mathbf{x}_{t-1} \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}), \quad (11)$$

we have

$$\mathbb{E}_{s\dots(t-1)} \left\| \mathbf{x}_t - \overline{\mathbf{x}}_t \mathbf{1}_n^\top \right\|^2 \leq (q + p\rho^2)^{(t-s)} \left\| \mathbf{x}_s - \overline{\mathbf{x}}_s \mathbf{1}_n^\top \right\|^2,$$

where s is a time stamp earlier than t, $\overline{\mathbf{x}}_t$ is the average value of vector $\mathbf{x}_t$, and W_t is a doubly stochastic matrix.

Remark. Lemma 1 indicates that the error between $\mathbf{x}_t$ and $\overline{\mathbf{x}}_t \mathbf{1}_n^\top$ can converge to 0 at a rate governed by $(q + p\rho^2)$. So the random gossip averaging can provably attain consensus if $(q + p\rho^2) < 1$.

Lemma 2. For any matrix $X_t \in \mathbb{R}^{N\times n}$ calculated by the following formulation

$$X_t = X_0 \underbrace{\otimes M_0 W_0 \otimes M_1 \otimes M_2 W_1 \otimes M_3 W_2 \otimes M_4 \dots}_{t \text{ mask matrices and } k \text{ gossip matrices}}, \quad (12)$$

for any multiplication order, we can rewrite

$$X_t = X_0 \bigsqcup_{i=0}^{t-1} M_i \prod_{i=0}^{k-1} W_i. \quad (13)$$

Proof. For any matrix $A \in \mathbb{R}^{N\times n}$, we define two matrices $Y = A \otimes MW$ and $Z = AW \otimes M$. Here $M^{(i,1)} = M^{(i,2)} = \dots = M^{(i,n)}$, and the i-th raw and j-th column element of Y and Z have the following relationship:

$$Y^{(i,j)} = \sum_{k=1}^{n} A^{(i,k)} M^{(i,k)} W^{(k,j)} = \sum_{k=1}^{n} A^{(i,k)} M^{(i,j)} W^{(k,j)} = \left(\sum_{k=1}^{n} A^{(i,k)} W^{(k,j)}\right) M^{(i,j)} = Z^{(i,j)},$$

which indicates $Y = Z$, i.e., $A \otimes MW = AW \otimes M$. It means that we can exchange the product order between M and W. Rearranging the order of (12), and putting all matrix products with W_i at the end of the equation, we can obtain (13). □

We refer to this following lemma in [14].

Lemma 3. Given two non-negative sequences $\{a_t\}_{t=1}^{\infty}$ and $\{b_t\}_{t=1}^{\infty}$ that satisfying

$$a_t = \sum_{s=1}^{t} \rho^{t-s} b_s$$

with $\rho \in [0, 1)$, we have

$$S_k := \sum_{t=1}^{k} a_t \leq \sum_{s=1}^{k} a_t \frac{b_s}{1-\rho}$$

$$D_k := \sum_{t=1}^{k} a_t^2 \leq \frac{1}{(1-\rho)^2} \sum_{s=1}^{k} b_s^2,$$

which has been proved in the appendix of [14].

8.2 Proof of Lemma 1

Proof. We refer to the proof in [28] [53], considering the evolution of a row vector $\mathbf{y}_t$ defined as follows

$$\begin{aligned}
\mathbf{y}_t =& \mathbf{x}_t - \overline{\mathbf{x}}_t \mathbf{1}_n^\top \\
\overset{(a)}{=}& \mathbf{x}_{t-1} \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}) \\
& - \overline{\mathbf{x}}_{t-1} \mathbf{1}_n^\top \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}) \\
=& (\mathbf{x}_{t-1} - \overline{\mathbf{x}}_{t-1} \mathbf{1}_n^\top) \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}) \\
=& \mathbf{y}_{t-1} \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}),
\end{aligned}$$

where (a) holds from

$$\begin{aligned}
\overline{\mathbf{x}}_t \mathbf{1}_n^\top =& \mathbf{x}_t \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\
=& \mathbf{x}_{t-1} \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}) \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\
=& \mathbf{x}_{t-1} \otimes \neg \mathbf{m}_{t-1} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} + \mathbf{x}_{t-1} \otimes \mathbf{m}_{t-1} W_{t-1} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\
\overset{(b)}{=}& \mathbf{x}_{t-1} \otimes \neg \mathbf{m}_{t-1} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} + \mathbf{x}_{t-1} \otimes \mathbf{m}_{t-1} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} W_{t-1} \\
\overset{\text{Lemma 2}}{=}& \overline{\mathbf{x}}_{t-1} \mathbf{1}_n^\top \otimes \neg \mathbf{m}_{t-1} + \mathbf{x}_{t-1} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \otimes \mathbf{m}_{t-1} W_{t-1} \\
=& \overline{\mathbf{x}}_{t-1} \mathbf{1}_n^\top \otimes (\neg \mathbf{m}_{t-1} + \mathbf{m}_{t-1} W_{t-1}),
\end{aligned}$$

where (b) holds from the property of doubly stochastic matrix. That is

$$\mathbf{x} W \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} = \mathbf{x} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} = \mathbf{x} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} W.$$

Then, we have

$$\begin{aligned}
& \mathbb{E}_{t-1} \|\mathbf{y}_t\|^2 = \mathbb{E}_{t-1} \mathbf{y}_t \mathbf{y}_t^\top \\
=& \mathbb{E}_{t-1} \left[(\mathbf{y}_{t-1} \otimes \neg \mathbf{m}_{t-1} + \mathbf{y}_{t-1} \otimes \mathbf{m}_{t-1} W_{t-1}) \right. \\
& \left. (\mathbf{y}_{t-1} \otimes \neg \mathbf{m}_{t-1} + \mathbf{y}_{t-1} \otimes \mathbf{m}_{t-1} W_{t-1})^\top \right] \\
=& \mathbb{E}_{t-1} \left[(\mathbf{y}_{t-1} \otimes \neg \mathbf{m}_{t-1})(\mathbf{y}_{t-1} \otimes \neg \mathbf{m}_{t-1})^\top \right. \\
& \left. + (\mathbf{y}_{t-1} \otimes \mathbf{m}_{t-1} W_{t-1})(\mathbf{y}_{t-1} \otimes \mathbf{m}_{t-1} W_{t-1})^\top \right] \\
=& q \mathbf{y}_{t-1} \mathbf{y}_{t-1}^\top + p \mathbb{E}_{t-1} \left[(\mathbf{y}_{t-1} \otimes W_{t-1})(\mathbf{y}_{t-1} \otimes W_{t-1})^\top \right] \\
=& q \|\mathbf{y}_{t-1}\|^2 + p \mathbf{y}_{t-1} \mathbb{E} \left[W_{t-1} W_{t-1}^\top \right] \mathbf{y}_{t-1}^\top
\end{aligned} \quad (14)$$

W_t is doubly stochastic, so $W_t^\top W_t$ and $\mathbb{E}\left[W_t^\top W_t\right]$ are both doubly stochastic. Because W_t is i.i.d., we can rewrite $\mathbb{E}\left[W_t^\top W_t\right]$ as $\mathbb{E}\left[W W^\top\right]$. Since $\mathbf{y}_t \perp \mathbf{1}_n$ [28], and $\mathbf{1}_n$ is the eigenvector corresponding to the largest eigenvalue 1 of $\mathbb{E}\left[W W^\top\right]$, we have [28]

$$\mathbf{y}_{t-1} \mathbb{E}\left[W W^\top\right] \mathbf{y}_{t-1}^\top \leq \rho^2 \|\mathbf{y}_{t-1}\|^2. \quad (15)$$

As $\mathbb{E}\left[WW^\top\right]=\mathbb{E}[W]$ [28], the second largest eigenvalue of $\mathbb{E}\left[WW^\top\right]$ should be equal to ρ. Repeatedly combining (14) and (15) we can obtain

$$\mathbb{E}_{W_s,W_{s+1},\dots W_{t-1}}\|\mathbf{y}_t\|^2\leq(q+p\rho^2)^{(t-s)}\|\mathbf{y}_s\|^2. \tag{16}$$

Substitute $\mathbf{y}_t=\mathbf{x}_t-\overline{\mathbf{x}}_t\mathbf{1}_n^\top$ into (16), we complete the proof. □

8.3 Analysis for GossipFL

Lemma 4. Under the assumptions defined in Section (4.3), if X_t is iteratively updated by Eq. (10), then we have

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)}-\overline{X_t}\|^2\leq\frac{2}{1-(q+p\rho^2)}\|X_0-\overline{X_0}\mathbf{1}_n^\top\|_F^2+\frac{2}{(1-(q+p\rho^2)^{\frac{1}{2}})^2}\sum_{t=1}^{T}\gamma_t^2\mathbb{E}\|G(X_t;\xi_t)\|_F^2, \tag{17}$$

where $\overline{X}$=$X\frac{\mathbf{1}_n}{n}$.

Proof. From Eq. (10), we have

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)}-\overline{X_t}\|^2\\
=&\sum_{i=1}^{n}\mathbb{E}\|X_t\mathbf{e}_n^{(i)}-X_t\frac{\mathbf{1}_n}{n}\|^2=\mathbb{E}\|X_t-X_t\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|_F^2\\
=&\mathbb{E}\|X_t(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|_F^2=\sum_{j=1}^{N}\mathbb{E}\|\mathbf{e}_N^{[i]}X_t(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
=&\sum_{j=1}^{N}\mathbb{E}\|\mathbf{e}_N^{[i]}(X_0\bigsqcup_{s=0}^{t-1}(\neg M_s+M_sW_s))(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\\
&-\mathbf{e}_N^{[i]}(\sum_{s=0}^{t-1}\gamma_sG(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg M_r+M_rW_r))(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
=&\sum_{j=1}^{N}\mathbb{E}\|\mathbf{x}_0^{[j]}\bigsqcup_{s=0}^{t-1}(\neg\mathbf{m}_s^{[j]}+\mathbf{m}_s^{[j]}W_s)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\\
&-\sum_{s=0}^{t-1}\gamma_sG^{[j]}(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
\leq&2\sum_{j=1}^{N}(\underbrace{\mathbb{E}\|\mathbf{x}_0^{[j]}\bigsqcup_{s=0}^{t-1}(\neg\mathbf{m}_s^{[j]}+\mathbf{m}_s^{[j]}W_s)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2}_{Q1}\\
&+\underbrace{\mathbb{E}\|\sum_{s=0}^{t-1}H_s^{[j]}\|^2}_{Q2}),
\end{aligned}\tag{18}$$

where

$$H_s^{[j]}=\gamma_sG^{[j]}(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}).$$

By using Lemma 1, we have

$$\begin{aligned}
Q1=&\mathbb{E}\|\mathbf{x}_0^{[j]}\bigsqcup_{s=0}^{t-1}(\neg\mathbf{m}_s^{[j]}+\mathbf{m}_s^{[j]}W_s)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
\leq&\|\mathbf{x}_0^{[j]}-\mathbf{x}_0^{[j]}\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|^2(q+p\rho^2)^t.
\end{aligned}\tag{19}$$

For $Q2$, we have

$$\begin{aligned}
\mathbb{E}\|\sum_{s=0}^{t-1}H_s^{[j]}\|^2=&\sum_{s=0}^{t-1}\mathbb{E}\|H_s^{[j]}\|^2+2\sum_{s<z}^{t-1}\mathbb{E}\left\langle H_s^{[j]},H_z^{[j]}\right\rangle\\
\leq&\sum_{s=0}^{t-1}\mathbb{E}\|H_s^{[j]}\|^2+2\sum_{s<z}^{t-1}\mathbb{E}\|H_s^{[j]}\|\|H_z^{[j]}\|.
\end{aligned}\tag{20}$$

We firstly bound $\mathbb{E}\|H_s^{[j]}\|^2$. Note that $\mathbb{E}\|H_s^{[j]}\|^2$ has the same formula with $Q1$. Thus, using Lemma 1, we have,

$$\begin{aligned}
&\mathbb{E}\|H_s^{[j]}\|^2\\
=&\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
=&\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)-\gamma_sG^{[j]}(X_s;\xi_s)\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|^2(q+p\rho^2)^{t-s-1}\\
\leq&\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|^2(q+p\rho^2)^{t-s-1}.
\end{aligned}\tag{21}$$

Then we bound $\mathbb{E}\|H_s^{[j]}\|\|H_z^{[j]}\|$, i.e.,

$$\begin{aligned}
&\mathbb{E}\|H_s^{[j]}\|\|H_z^{[j]}\|\\
=&\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|\\
&\cdot\|\gamma_zG^{[j]}(X_z;\xi_z)\bigsqcup_{r=z+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|\\
\leq&\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|(q+p\rho^2)^{(t-s-1)/2}\\
&\cdot\|\gamma_zG^{[j]}(X_z;\xi_z)\|(q+p\rho^2)^{(t-z-1)/2}.
\end{aligned}\tag{22}$$

Combining (20), (21) and (22), we can bound $Q2$ as

$$\begin{aligned}
&\mathbb{E}\|\sum_{s=0}^{t-1}\gamma_sG^{[j]}(X_s;\xi_s)\bigsqcup_{r=s+1}^{t-1}(\neg\mathbf{m}_r^{[j]}+\mathbf{m}_r^{[j]}W_r)(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n})\|^2\\
\leq&\sum_{s=0}^{t-1}\mathbb{E}\|H_s^{[j]}\|^2+2\sum_{s<z}^{t-1}\mathbb{E}\|H_s^{[j]}\|\|H_z^{[j]}\|\\
\leq&\sum_{s=0}^{t-1}\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|^2(q+p\rho^2)^{t-s-1}\\
&+2\sum_{s<z}^{t-1}\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|\cdot\|\gamma_zG^{[j]}(X_z;\xi_z)\|\\
&\cdot(q+p\rho^2)^{\frac{1}{2}(t-s-1)}(q+p\rho^2)^{\frac{1}{2}(t-z-1)}\\
=&\left(\sum_{s=0}^{t-1}\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|(q+p\rho^2)^{\frac{1}{2}(t-s-1)}\right)^2.
\end{aligned}\tag{23}$$

Combining (18) , (19) and (23), we have

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)}-\overline{X_t}\|^2\\
\leq&2\sum_{j=1}^{N}\left(\|\mathbf{x}_0^{[j]}-\mathbf{x}_0^{[j]}\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|^2(q+p\rho^2)^t\right)\\
&+2\sum_{j=1}^{N}\left(\sum_{s=0}^{t-1}\mathbb{E}\|\gamma_sG^{[j]}(X_s;\xi_s)\|(q+p\rho^2)^{\frac{1}{2}(t-s-1)}\right)^2,
\end{aligned}\tag{24}$$

where $\left(\sum_{s=0}^{t-1}\|\gamma_sG^{[j]}(X_s;\xi_s)\|(q+p\rho^2)^{\frac{1}{2}(t-s-1)}\right)^2$ has the same structure with the sum of geometric sequence and

Lemma (3). Therefore, we sum (24) from $t = 1$ to $t = T$ to obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 \\
\leq& 2\sum_{j=1}^{N}\sum_{t=1}^{T}\left(\|\mathbf{x}_0^{[j]} - \mathbf{x}_0^{[j]}\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|^2(q+p\rho^2)^t\right) \\
&+ 2\sum_{j=1}^{N}\sum_{t=1}^{T}\left(\sum_{s=0}^{t-1}\mathbb{E}\|\gamma_s G^{[j]}(X_s;\xi_s)\|(q+p\rho^2)^{\frac{1}{2}(t-s-1)}\right)^2 \\
\leq& \frac{2}{1-(q+p\rho^2)}\sum_{j=1}^{N}\left(\|\mathbf{x}_0^{[j]} - \mathbf{x}_0^{[j]}\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\|^2\right) \\
&+ \frac{2}{(1-(q+p\rho^2)^{\frac{1}{2}})^2}\sum_{j=1}^{N}\sum_{t=1}^{T}\mathbb{E}\|\gamma_s G^{[j]}(X_s;\xi_s)\|^2 \\
\leq& \frac{2}{1-(q+p\rho^2)}\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2 \\
&+ \frac{2}{(1-(q+p\rho^2)^{\frac{1}{2}})^2}\sum_{t=1}^{T}\gamma_t^2\mathbb{E}\|G(X_t;\xi_t)\|_F^2, \qquad (25)
\end{aligned}$$

which completes the proof of Lemma 4. □

Note that if we make all clients have the same initial parameters (i.e., $\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|^2 = 0$), then the consensus is only impacted by the gradients, which indicates that the random gossip averaging with sparsified communication can attain consensus.

Now we begin to prove the convergence of our algorithm. We firstly introduce some lemmas.

Lemma 5. Following the assumptions defined in Section 4.3, we have

$$\begin{aligned}
&\frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2 + \left(\frac{\gamma_t}{2} - \frac{L\gamma_t^2}{2}\right)\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\leq& \mathbb{E}f(\overline{X}_t) - \mathbb{E}f(\overline{X}_{t+1}) + \frac{L^2\gamma_t}{2n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + \frac{L\gamma_t^2\sigma^2}{2n}.s
\end{aligned}$$

Proof. From the updating rule, we have

$$\begin{aligned}
\overline{X}_{t+1} &= X_{t+1}\frac{\mathbf{1}_n}{n} \\
&= X_t \otimes \neg M_t\frac{\mathbf{1}_n}{n} + X_t \otimes M_t W_t\frac{\mathbf{1}_n}{n} - \gamma_t G(X_t;\xi_t)\frac{\mathbf{1}_n}{n} \\
&= X_t \otimes \neg M_t\frac{\mathbf{1}_n}{n} + X_t \otimes M_t\frac{\mathbf{1}_n}{n} - \gamma_t\overline{G}(X_t;\xi_t) \\
&= \overline{X}_t - \gamma_t\overline{G}(X_t;\xi_t).
\end{aligned}$$

According to the Lipschitzian condition for the objective function f_i and f, we have

$$\begin{aligned}
&\mathbb{E}f(\overline{X}_{t+1}) \\
\leq& \mathbb{E}f(\overline{X}_t) + \mathbb{E}\left\langle\nabla f(\overline{X}_t), \gamma_t\overline{G}(X_t;\xi_t)\right\rangle + \frac{L}{2}\mathbb{E}\| - \gamma_t\overline{G}(X_t;\xi_t)\|^2 \\
=& \mathbb{E}f(\overline{X}_t) + \mathbb{E}\left\langle\nabla f(\overline{X}_t), -\gamma_t\mathbb{E}_{\xi_t}\overline{G}(X_t;\xi_t)\right\rangle \\
&+ \frac{L\gamma_t^2}{2}\mathbb{E}\|(\overline{G}(X_t;\xi_t) - \overline{\nabla f}(X_t)) + \overline{\nabla f}(X_t)\|^2 \\
=& \mathbb{E}f(\overline{X}_t) - \gamma_t\mathbb{E}\left\langle\nabla f(\overline{X}_t), \overline{\nabla f}(X_t)\right\rangle \\
&+ \frac{L\gamma_t^2}{2}\mathbb{E}\|\overline{G}(X_t;\xi_t) - \overline{\nabla f}(X_t)\|^2 + \frac{L\gamma_t^2}{2}\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
=& \mathbb{E}f(\overline{X}_t) - \gamma_t\mathbb{E}\left\langle\nabla f(\overline{X}_t), \overline{\nabla f}(X_t)\right\rangle \\
&+ \frac{L\gamma_t^2}{2n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)})\|^2 + \frac{L\gamma_t^2}{2}\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\leq& \mathbb{E}f(\overline{X}_t) - \frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2 - \frac{\gamma_t}{2}\mathbb{E}\|\overline{\nabla f}(X_t))\|^2 \\
&+ \frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t) - \overline{\nabla f}(X_t))\|^2 + \frac{L\gamma_t^2\sigma^2}{2n} + \frac{L\gamma_t^2}{2}\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
=& \mathbb{E}f(\overline{X}_t) - \frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2 - \left(\frac{\gamma_t}{2} - \frac{L\gamma_t^2}{2}\right)\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
&+ \frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t) - \overline{\nabla f}(X_t)\|^2 + \frac{L\gamma_t^2\sigma^2}{2n}. \qquad (26)
\end{aligned}$$

We can bound $\mathbb{E}\|\nabla f(\overline{X}_t) - \overline{\nabla f}(X_t))\|^2$ as

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\overline{X}_t) - \overline{\nabla f}(X_t))\|^2 \\
=& \frac{1}{n^2}\mathbb{E}\|\sum_{i=1}^{n}\left(\nabla f_i(\overline{X}_t) - \nabla f_i(\mathbf{x}_t^{(i)})\right)\|^2 \\
\leq& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\overline{X}_t) - \nabla f_i(\mathbf{x}_t^{(i)})\|^2 \\
\leq& \frac{L^2}{n}\mathbb{E}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2, \qquad (27)
\end{aligned}$$

Combining (26) and (27) together and rearranging, we have

$$\begin{aligned}
&\frac{\gamma_t}{2}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2 + \left(\frac{\gamma_t}{2} - \frac{L\gamma_t^2}{2}\right)\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\leq& \mathbb{E}f(\overline{X}_t) - \mathbb{E}f(\overline{X}_{t+1}) + \frac{L^2\gamma_t}{2n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + \frac{L\gamma_t^2\sigma^2}{2n}
\end{aligned}$$

which completes the proof. □

Lemma 6. Under the assumptions defined in Section 4.3, we can bound $\mathbb{E}\|G(X_t;\xi_t)\|_F^2$ as follows

$$\begin{aligned}
\mathbb{E}\|G(X_t;\xi_t)\|_F^2 \leq& n\sigma^2 + 3L^2\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 \\
&+ 3n\zeta^2 + 3n\mathbb{E}\|\nabla f(\overline{X}_t)\|^2.
\end{aligned}$$

Proof. Notice that

$$\begin{aligned}
&\mathbb{E}\|G(X_t;\xi_t)\|_F^2 = \sum_{i=1}^n \mathbb{E}\|\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)})\|_F^2 \\
=&\sum_{i=1}^n \mathbb{E}\| \left(\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)})\right) + \nabla f_i(\mathbf{x}_t^{(i)})\|^2 \\
=&\sum_{i=1}^n \mathbb{E}\|\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)})\|^2 + \sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}_t^{(i)})\|^2 \\
&+ 2\sum_{i=1}^n \mathbb{E}\left\langle \mathbb{E}_{\xi_t}\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)}), \nabla f_i(\mathbf{x}_t^{(i)})\right\rangle \\
=&\sum_{i=1}^n \mathbb{E}\|\nabla F_i(\mathbf{x}_t^{(i)};\xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)})\|^2 + \sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}_t^{(i)})\|^2 \\
\le& n\sigma^2 + \sum_{i=1}^n \mathbb{E}\|(\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\overline{X}_t)) \\
&+ \left(\nabla f_i(\overline{X}_t) - \nabla f(\overline{X}_t)\right) + \nabla f(\overline{X}_t)\|^2 \\
\le& n\sigma^2 + 3\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\overline{X}_t)\|^2 \\
&+ 3\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\overline{X}_t) - \nabla f(\overline{X}_t)\|^2 + 3\sum_{i=1}^n \mathbb{E}\|\nabla f(\overline{X}_t)\|^2 \\
\le& n\sigma^2 + 3L^2\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + 3n\zeta^2 + 3n\mathbb{E}\|\nabla f(\overline{X}_t)\|^2,
\end{aligned}$$

which completes the proof. □

Next, we use Lemma 6 to substitute $\|G(X_t;\xi_t)\|^2$ in Lemma 4.

Lemma 7. Under assumptions defined in Section 4.3, we have

$$\begin{aligned}
\sum_{t=1}^T (1-3D_1L^2\gamma_t^2)\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 \le D_1 n(\sigma^2+3\zeta^2)\sum_{t=1}^T \gamma_t^2 \\
+3nD_1\sum_{t=1}^T \gamma_t^2\|\nabla f(\overline{X}_t)\|^2 + D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2,
\end{aligned}$$

where $D_1 = \frac{2}{(1-(q+p\rho^2)^{\frac{1}{2}})^2}$ and $D_2 = \frac{2}{1-(q+p\rho^2)}$.

Proof. Combining (25) and using Lemma 6, we have

$$\begin{aligned}
&\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 \\
\le& D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2 + D_1\sum_{t=1}^T \gamma_t^2\|G(X_t;\xi_t)\|_F^2 \\
\le& D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2 + D_1 n(\sigma^2+3\zeta^2)\sum_{t=1}^T \gamma_t^2 \\
&+ 3D_1L^2\sum_{t=1}^T \gamma_t^2\sum_{i=1}^n \|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + 3D_1 n\sum_{t=1}^T \gamma_t^2\mathbb{E}\|\nabla f(\overline{X}_t)\|^2.
\end{aligned}$$

Rearranging the above equation, it yields

$$\begin{aligned}
\sum_{t=1}^T (1-3D_1L^2\gamma_t^2)\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 \le D_1 n(\sigma^2+3\zeta^2)\sum_{t=1}^T \gamma_t^2 \\
+3nD_1\sum_{t=1}^T \gamma_t^2\|\nabla f(\overline{X}_t)\|^2 + D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2.
\end{aligned}$$

If $1-3D_1L^2\gamma_t^2 > 0$, then $\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2$ is bounded. □

8.3.1 Proof of Theorem 1

Proof. From Lemma 5, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\overline{X}_t)\|^2 + (1-L\gamma_t)\,\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\le&\frac{2}{\gamma_t}\left(\mathbb{E}f(\overline{X}_t) - f^* - (\mathbb{E}f(\overline{X}_{t+1}) - f^*)\right) \\
&+ \frac{L^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + \frac{L\gamma_t\sigma^2}{n}.
\end{aligned} \tag{28}$$

According to Lemma 7, if we fix γ_t to satisfy $1-6D_1L^2\gamma > 0$ and sum both sides of (28), we obtain

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\|\nabla f(\overline{X}_t)\|^2 + \sum_{t=1}^T (1-L\gamma)\,\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\le&\frac{2}{\gamma}\left(f(X_0) - f^*\right) + \frac{L^2}{n}\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_t^{(i)} - \overline{X_t}\|^2 + \frac{LT\gamma\sigma^2}{n} \\
\le&\frac{2}{\gamma}\left(f(X_0) - f^*\right) + \frac{L^2}{n}\left(\frac{D_1 n(\sigma^2+3\zeta^2)T\gamma^2}{1-3D_1L^2\gamma^2}\right. \\
&\left. + \frac{3nD_1\gamma^2}{1-3D_1L^2\gamma^2}\sum_{t=1}^T \|\nabla f(\overline{X}_t)\|^2 + \frac{D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2}{1-3D_1L^2\gamma^2}\right) \\
&+ \frac{LT\gamma\sigma^2}{n}.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{1-6D_1L^2\gamma^2}{1-3D_1L^2\gamma^2}\sum_{t=1}^T \mathbb{E}\|\nabla f(\overline{X}_t)\|^2 + \sum_{t=1}^T (1-L\gamma)\,\mathbb{E}\|\overline{\nabla f}(X_t)\|^2 \\
\le&\frac{2}{\gamma}\left(f(X_0) - f^*\right) + \left(\frac{L^2D_1T\gamma^2}{1-3D_1L^2\gamma^2} + \frac{L\gamma T}{n}\right)\sigma^2 \\
&+ \frac{3L^2D_1T\gamma^2\zeta^2}{1-3D_1L^2\gamma^2} + \frac{L^2D_2\|X_0 - \overline{X_0}\mathbf{1}_n^\top\|_F^2}{n(1-3D_1L^2\gamma^2)},
\end{aligned} \tag{29}$$

which completes the proof. □

8.3.2 Proof of Corollary 1

Proof. Setting $\gamma = \frac{1}{2\sqrt{3D_1}L + \frac{\sigma}{\sqrt{n}}\sqrt{T} + \zeta T^{\frac{1}{3}}}$, it yields

$$3D_1L^2\gamma^2 \le \frac{1}{4}, \tag{30}$$

$$\frac{1-6D_1L^2\gamma^2}{1-3D_1L^2\gamma^2} \ge \frac{2}{3}, \tag{31}$$

$$1 - L\gamma > 0. \tag{32}$$

Then we can remove the $(1-L\gamma)\mathbb{E}\|\overline{\nabla f}(X_t)\|^2$ on the left hand side of Eq. (7) and substitute $\frac{1-6D_1L^2\gamma^2}{1-3D_1L^2\gamma^2}$ with $\frac{2}{3}$, so we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2\\
\leq&\frac{3}{\gamma}\left(f(X_0)-f^*\right)+\left(2L^2D_1\gamma^2+\frac{3L\gamma}{2n}\right)\sigma^2\\
&+6L^2D_1\gamma^2\zeta^2+\frac{2L^2D_2\|X_0-\overline{X_0}\mathbf{1}_n^\top\|_F^2}{nT}\\
\leq&(\frac{6\sqrt{3D_1}L}{T}+\frac{3\sigma}{\sqrt{nT}}+\frac{3\zeta}{T^{\frac{2}{3}}})\left(f(X_0)-f^*\right)\\
&+\frac{2L^2nD_1}{T}+\frac{3L\sigma}{2\sqrt{nT}}\\
&+\frac{6L^2D_1}{T^{\frac{2}{3}}}+\frac{2L^2D_2\|X_0-\overline{X_0}\mathbf{1}_n^\top\|_F^2}{nT},
\end{aligned}
$$

which means

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{X}_t)\|^2\lesssim&\frac{\sigma}{\sqrt{nT}}\\
&+(\zeta+D_1)\frac{1}{T^{\frac{2}{3}}}+(nD_1+\sqrt{D_1}+\frac{D_2}{n})\frac{1}{T},
\end{aligned}
\tag{33}
$$

which completes the proof. □

8.3.3 Proof of Theorem 2

Proof. Each column vector in matrix X_t represents the local model of one client. According the definition of consensus, each column vector of X_t should be the same vector. Under the assumptions defined in Section 4.3, combining Lemma (7) and Eq. (30)(33), and setting $\gamma=\frac{1}{2\sqrt{3D_1}L+\frac{\sigma}{\sqrt{n}}\sqrt{T}+\zeta T^{\frac{1}{3}}}$, we can obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_t^{(i)}-\overline{X_t}\|^2\leq D_1n(\sigma^2+3\zeta^2)\gamma^2\\
&+\frac{3nD_1}{T}\sum_{t=1}^{T}\gamma^2\|\nabla f(\overline{X}_t)\|^2+\frac{D_2}{T}\|X_0-\overline{X_0}\mathbf{1}_n^\top\|_F^2\\
\lesssim&\frac{D_1n^2\sigma}{T^2}+\frac{D_1n}{T^{\frac{2}{3}}}+\frac{D_2}{T}+\frac{\sqrt{D_1}n+n^2D_1}{T},
\end{aligned}
$$

which concludes the proof.

□