A APPENDIX

A.1 PROOF OF LEMMA 4.2.

Proof. Notice that for any scalar $w \in \mathbb{R}^n$ and $l \leq w \leq u$, by the definition of $\mathcal{M}(.; l, u)$, one can verify that $\mathbb{E}[\mathcal{M}(w; l, u)] = w$, therefore $\mathbb{E}[\mathcal{M}(\boldsymbol{w}; l, u)] = \boldsymbol{w}$. Furthermore,

$$
\begin{aligned}
\mathbb{E}[\|\mathcal{M}(\boldsymbol{w}; l, u) - \boldsymbol{w}\|^2] &= \sum_{j=1}^{d} \mathbb{E}[([\mathcal{M}(\boldsymbol{w}; l, u)]_j - [\boldsymbol{w}]_j)^2] \\
&= \sum_{j=1}^{d} \left(\mathbb{E}[[\mathcal{M}(\boldsymbol{w}; l, u)]_j^2] - [\boldsymbol{w}]_j^2\right) \\
&= \sum_{j=1}^{d} \left((u + l)[\boldsymbol{w}]_j - lu - [\boldsymbol{w}]_j^2\right) \\
&= \sum_{j=1}^{d} \left[\left(\frac{u-l}{2}\right)^2 - \left([\boldsymbol{w}]_j - \left(\frac{u+l}{2}\right)\right)^2\right] \\
&= d\left(\frac{u-l}{2}\right)^2 - \sum_{j=1}^{d} \left([\boldsymbol{w}]_j - \left(\frac{u+l}{2}\right)\right)^2 \leq \frac{d(u-l)^2}{4}.
\end{aligned}
$$

Moreover,

$$
d\left(\frac{u-l}{2}\right)^2 - \sum_{j=1}^{d} \left([\boldsymbol{w}]_j - \left(\frac{u+l}{2}\right)\right)^2 =
$$

□

A.2 PROOF OF THEOREM 4.2.

Before proving the Theorem 4.2, a standard probability bound is required.

Lemma A.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d continuous random variables, whose support is on $\mathbb{R}$. Then for any $b \in \mathbb{R}$*

$$
\mathbb{P}\left(\max_{i \in [n]} X_i \geq b\right) \leq n\mathbb{P}(X_i \geq b).
$$

Proof.

$$
\begin{aligned}
\mathbb{P}\left(\max_{i \in [n]} X_i \geq b\right) &= 1 - \mathbb{P}(\max_{i \in [n]} X_i \leq b) = 1 - \mathbb{P}(X_1 \leq b, \cdots, X_n \leq b) \\
&= 1 - \prod_{j=1}^{n} \mathbb{P}(X_j \leq b) = 1 - (1 - \mathbb{P}(X_j \geq b))^n \text{ for any } j
\end{aligned}
$$

Let $f(t) := 1 - nt - (1-t)^n$ for $t \in [0, 1]$. As $f'(t) \leq 0$ and $f(0) = 0$, $f(t) \leq 0$ for all $t \in [0, 1]$. Therefore, $1 - (1-t)^n \leq nt$. Take $t = \mathbb{P}(X_j \geq b)$, we arrive

$$
\mathbb{P}\left(\max_{i \in [n]} X_i \geq b\right) \leq n\mathbb{P}(X_i \geq b).
$$

□

Now we are ready to prove Theorem 4.2.

Proof. For any $j \in [d]$, $|[\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j| \leq u^t - l^t$ and $\mathbf{Var}\left([\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j\right) \leq (u^t - l^t)^2/4$ due to Lemma 4.1. By Bernstein inequality, for any $j \in [d]$ and $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}\left(\left|[\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j\right| \geq \epsilon\right) &\leq 2\exp\left(-\frac{K\epsilon^2}{2\frac{1}{K}\sum_{i\in S_t}\mathbf{Var}\left([\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j\right) + \frac{2}{3}\epsilon(u^t - l^t)}\right) \\
&\leq 2\exp\left(-\frac{K\epsilon^2}{\frac{(u^t-l^t)^2}{2} + \frac{2}{3}\epsilon(u^t - l^t)}\right)
\end{aligned}$$

By Lemma A.1,

$$\mathbb{P}\left(\max_{j\in[d]}\left|[\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j\right| \geq \epsilon\right) \leq 2d\exp\left(-\frac{K\epsilon^2}{\frac{(u^t-l^t)^2}{2} + \frac{2}{3}\epsilon(u^t - l^t)}\right)$$

Therefore, for any $\beta > 0$, there exists $\epsilon = \mathcal{O}\left(\frac{(u^t-l^t)\sqrt{\log\frac{2d}{\beta}}}{\sqrt{K}}\right)$ such that $\mathbb{P}\left(\max_{j\in[d]}|[\bar{\boldsymbol{w}}^t]_j - [\bar{\boldsymbol{w}}^t_{\mathcal{M}}]_j| \leq \epsilon\right)$ holds with probability at least $1-\beta$. □

A.3 Proof of Theorem 4.3

Proof. Due to the weight discretization mechanism, the adversary can only return u or l for each coordinate of the model weight. In order to not return the correct information, the adversary could choose to return the opposite feedback to attack the model, i.e., return u if the original return is l, and vice versa. Therefore, we denote, for any $l \leq w \leq u$,

$$\mathcal{M}_{\text{adv}}(w) = \begin{cases} l, & \text{w.p.} \frac{w-l}{h-l} \\ h, & \text{w.p.} \frac{h-w}{h-l} \end{cases}$$

Under the scenario that there are F attackers, for any $j \in [d]$,

$$\begin{aligned}
[\mathbb{E}[\bar{\boldsymbol{w}}_{\mathcal{M}}]]_j &= \mathbb{E}\left[\frac{1}{N}\left(\sum_{i=1}^{N-F}[\mathcal{M}(\boldsymbol{w}_i)]_j + \sum_{i=N-F+1}[\mathcal{M}_{\text{adv}}(\boldsymbol{w}_i)]_j\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N-F}[\boldsymbol{w}_i]_j + \frac{1}{N}\sum_{i=N-F+1}^{N}((h+l) - [\boldsymbol{w}_i]_j) && (4) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N-F}[\boldsymbol{w}_i]_j - \sum_{i=N-F+1}^{N}[\boldsymbol{w}_i]_j\right) + (h+l)\frac{F}{N}. && (5)
\end{aligned}$$

□

A.4 Proof of Theorem 4.4

Proof. The proof is inspired by the analysis in Theorem 4 of (Li et al., 2020).

To proceed with the analysis, we first introduce some notations. At the tth round, for the all $i \in \mathcal{S}'_t$, define $\tilde{\boldsymbol{w}}^{t+1} = \boldsymbol{w}^t + \frac{1}{|\mathcal{S}'_t|}\sum_{i\in\mathcal{S}'_t}\left(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right)$, $\bar{\boldsymbol{w}}^{t+1} = \boldsymbol{w}^t + \sum_{i\in[N]}p_i(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t)$, and $\hat{\boldsymbol{w}}_i^{t+1} = \arg\min_{\boldsymbol{w}} h_i(\boldsymbol{w};\boldsymbol{w}^t) := F_i(\boldsymbol{w}) + \frac{\mu}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^t\right\|^2$. $\tilde{\boldsymbol{w}}^{t+1}$ is the ghost global model as if the discretization mechanism is not applied to the local model weights; $\bar{\boldsymbol{w}}^{t+1}$ is another ghost global model as if all clients participate in the tth round training and no discretization mechanism is applied ; $\hat{\boldsymbol{w}}_i^{t+1}$ is the exact minimizer of the strongly convex function $h_i(\boldsymbol{w})$. These points reference points are crucial for the analysis. Define the gradient residual $e_i^{t+1} = \nabla F_i(\boldsymbol{w}_i^{t+1}) + \mu(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t)$, then $\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t = -\frac{1}{\mu}\nabla F_i(\boldsymbol{w}_i^{t+1}) + \frac{1}{\mu}e_i^{t+1}$. Therefore,

$$\bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t = \sum_{i\in[n]}p_i(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t) = -\frac{1}{\mu}\sum_{i\in[n]}p_i\nabla F_i(\boldsymbol{w}_i^{t+1}) + \frac{1}{\mu}\sum_{i\in[n]}p_i e_i^{t+1}. \tag{6}$$

Since μ is chosen to satisfy $\mu > \lambda_{\min}$, then $h_i(\boldsymbol{w}; \boldsymbol{w}^t)$ is $\bar{\mu}$-strongly convex. By the strong convexity of h_i,

$$\begin{aligned}
\left\|\boldsymbol{w}_i^{t+1} - \hat{\boldsymbol{w}}_i^{t+1}\right\|^2 &\leq \frac{1}{\bar{\mu}}(\boldsymbol{w}_i^{t+1} - \hat{\boldsymbol{w}}_i^{t+1})^\top(\nabla h_i(\boldsymbol{w}_i^{t+1}) - \nabla h_i(\hat{\boldsymbol{w}}_i^{t+1})) \\
&\leq \frac{1}{\bar{\mu}}\left\|\boldsymbol{w}_i^{k+1} - \hat{\boldsymbol{w}}_i^{t+1}\right\|\left\|\nabla h_i(\boldsymbol{w}_i^{t+1}) - \nabla h_i(\hat{\boldsymbol{w}}_i^{t+1})\right\|,
\end{aligned}$$

which, together with the fact that $\hat{\boldsymbol{w}}_i^{t+1}$ is the minimizer of $h_i(\boldsymbol{w})$, implies

$$\begin{aligned}
\left\|\boldsymbol{w}_i^{t+1} - \hat{\boldsymbol{w}}_i^{t+1}\right\| &\leq \frac{1}{\bar{\mu}}\left\|\nabla h_i(\boldsymbol{w}_i^{t+1}) - \nabla h_i(\hat{\boldsymbol{w}}_i^{t+1}))\right\| \\
&= \frac{1}{\bar{\mu}}\left\|\nabla h_i(\hat{\boldsymbol{w}}_i^{t+1}))\right\| = \frac{1}{\bar{\mu}}\left\|\nabla F_i(\boldsymbol{w}_i^{t+1}) + \mu(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t)\right\| \\
&\leq \frac{\gamma}{\bar{\mu}}\left\|\nabla F_i(\boldsymbol{w}^t)\right\| \qquad (7)
\end{aligned}$$

Again use the same analysis, one has $\left\|\hat{\boldsymbol{w}}_i^{t+1} - \boldsymbol{w}^t\right\| \leq \frac{1}{\bar{\mu}}\nabla F_i(\boldsymbol{w}^t)$. Therefore, together with Eq. 7,

$$\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right\| \leq \left\|\boldsymbol{w}_i^{t+1} - \hat{\boldsymbol{w}}_i^{t+1}\right\| + \left\|\hat{\boldsymbol{w}}_i^{t+1} - \boldsymbol{w}^t\right\| \leq \frac{1+\gamma}{\bar{\mu}}\left\|\nabla F_i(\boldsymbol{w}^t)\right\|. \qquad (8)$$

Therefore, one can bound the distance from the ghost global model $\bar{\boldsymbol{w}}^{t+1}$ to the current global weight as

$$\begin{aligned}
\left\|\bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t\right\| &= \left\|\sum_{i\in[N]} p_i(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t)\right\| \leq \sum_{i\in[N]} p_i\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right\| \\
&\leq \frac{1+\gamma}{\bar{\mu}}\sum_{i\in[N]} p_i\left\|\nabla F_i(\boldsymbol{w}^t)\right\| \qquad \text{(by Eq. 8)} \\
&\leq \frac{1+\gamma}{\bar{\mu}}\sqrt{\sum_{i\in[N]} p_i\left\|\nabla F_i(\boldsymbol{w}^t)\right\|^2} \qquad \text{(Jensen' Inequality)} \\
&= \frac{1+\gamma}{\bar{\mu}}\sqrt{\mathbb{E}_i[\left\|\nabla F_i(\boldsymbol{w}^t)\right\|^2]} \\
&\leq \frac{B(1+\gamma)}{\bar{\mu}}\left\|\nabla f(\boldsymbol{w}^t)\right\| \qquad \text{(by Assumption 4.1 (3))} \qquad (9)
\end{aligned}$$

Note that

$$\begin{aligned}
\left\|\sum_{i\in[n]} p_i\left(\nabla F_i(\boldsymbol{w}_i^{t+1}) - e_i^{t+1} - \nabla F_i(\boldsymbol{w}^t)\right)\right\| &\leq \sum_{i\in[n]} p_i\left(\left\|\nabla F_i(\boldsymbol{w}_i^{t+1}) - \nabla F_i(\boldsymbol{w}^t)\right\| + \left\|e_i^{t+1}\right\|\right) \\
&\leq \sum_{i\in[n]} p_i\left(L\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right\| + \left\|e_i^{t+1}\right\|\right) \\
&\overset{Eq.\ 8}{\leq} \left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma\right)\sum_{i\in[n]} p_i\left\|\nabla F_i(\boldsymbol{w}^t)\right\| \\
&= \left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma\right)\mathbb{E}_i[\nabla F_i(\boldsymbol{w}^t)] \\
&\overset{Eq.\ 8}{\leq} B\left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma\right)\left\|\nabla f(\boldsymbol{w}^t)\right\|. \qquad (10)
\end{aligned}$$

By Assumption 4.1 (1), one has

$$
\begin{aligned}
f(\bar{\boldsymbol{w}}^{t+1}) &\le f(\boldsymbol{w}^t) + \nabla f(\boldsymbol{w}^t)^\top (\bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t) + \frac{L}{2} \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
&\overset{Eq.\ 6}{\le} f(\boldsymbol{w}^t) + \nabla f(\boldsymbol{w}^t)^\top \left(-\frac{1}{\mu} \sum_{i\in[n]} p_i \nabla F_i(\boldsymbol{w}_i^{t+1}) + \frac{1}{\mu} \sum_{i\in[n]} p_i e_i^{t+1} \right) + \frac{L}{2} \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
&= f(\boldsymbol{w}^t) + \nabla f(\boldsymbol{w}^t)^\top \left(-\frac{1}{\mu} \sum_{i\in[n]} p_i \left(\nabla F_i(\boldsymbol{w}_i^{t+1}) - e_i^{t+1} - \nabla F_i(\boldsymbol{w}^t) \right) - \frac{1}{\mu} \nabla f(\boldsymbol{w}^t) \right) \\
&\quad + \frac{L}{2} \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
&\le f(\boldsymbol{w}^t) - \frac{1}{\mu} \left\| f(\boldsymbol{w}^t) \right\|^2 - \frac{1}{\mu} f(\boldsymbol{w}^t)^\top \left(\sum_{i\in[n]} p_i \left(\nabla F_i(\boldsymbol{w}_i^{t+1}) - e_i^{t+1} - \nabla F_i(\boldsymbol{w}^t) \right) \right) \\
&\quad + \frac{L}{2} \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
&\le f(\boldsymbol{w}^t) - \frac{1}{\mu} \left\| f(\boldsymbol{w}^t) \right\|^2 + \frac{1}{\mu} \left\| f(\boldsymbol{w}^t) \right\| \left\| \sum_{i\in[n]} p_i \left(\nabla F_i(\boldsymbol{w}_i^{t+1}) - e_i^{t+1} - \nabla F_i(\boldsymbol{w}^t) \right) \right\| \\
&\quad + \frac{L}{2} \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
&\overset{Eq.\ 10, Eq.\ 9}{\le} f(\boldsymbol{w}^t) - \frac{1}{\mu} \left\| f(\boldsymbol{w}^t) \right\|^2 + \frac{B}{\mu} \left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma \right) \left\| \nabla f(\boldsymbol{w}^t) \right\|^2 + \frac{L}{2} (\frac{B(1+\gamma)}{\bar{\mu}})2 \left\| \nabla f(\boldsymbol{w}^t) \right\|^2 \quad (11) \\
&= f(\boldsymbol{w}^t) - \left(\frac{1-\gamma B}{\mu} - \frac{LB(1+\gamma)}{\mu\bar{\mu}} - \frac{L(1+\gamma)^2 B^2}{2\bar{\mu}^2} \right) \left\| \nabla f(\boldsymbol{w}^t) \right\|^2 \quad (12)
\end{aligned}
$$

By mean-value theorem and triangular inequality, for some $\alpha \in [0, 1]$

$$
\begin{aligned}
f(\tilde{\boldsymbol{w}}^{t+1}) &\le f(\bar{\boldsymbol{w}}^{t+1}) + \left\| \nabla f(\alpha \tilde{\boldsymbol{w}}^{t+1} + (1-\alpha)\bar{\boldsymbol{w}}^{t+1}) \right\| \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\le f(\bar{\boldsymbol{w}}^{t+1}) + \left(\left\| \nabla f(\alpha \tilde{\boldsymbol{w}}^{t+1} + (1-\alpha)\bar{\boldsymbol{w}}^{t+1}) - \nabla f(\boldsymbol{w}^t) \right\| + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\le f(\bar{\boldsymbol{w}}^{t+1}) + \left(L \left\| \alpha \tilde{\boldsymbol{w}}^{t+1} + (1-\alpha)\bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\le f(\bar{\boldsymbol{w}}^{t+1}) + \left(L(\left\| \tilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| + \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\|) + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\|
\end{aligned}
\quad (13)
$$

Taking expectation with respect to the random index set $\mathcal{S}'_t$, one gets

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}'_t}[f(\tilde{\boldsymbol{w}}^{t+1})] &\le f(\bar{\boldsymbol{w}}^{t+1}) + \left(L \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \mathbb{E}_{\mathcal{S}'_t} \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\quad + L\mathbb{E}_{\mathcal{S}'_t}[\left\| \tilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\|] \\
&\le f(\bar{\boldsymbol{w}}^{t+1}) + \left(L \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \mathbb{E}_{\mathcal{S}'_t} \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\quad + L\mathbb{E}_{\mathcal{S}'_t}[\left(\left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| + \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| \right) \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\|] \\
&= f(\bar{\boldsymbol{w}}^{t+1}) + \left(2L \left\| \bar{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^t \right\| + \left\| \nabla f(\boldsymbol{w}^t) \right\| \right) \mathbb{E}_{\mathcal{S}'_t} \left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\| \\
&\quad + L\mathbb{E}_{\mathcal{S}'_t}[\left\| \tilde{\boldsymbol{w}}^{t+1} - \bar{\boldsymbol{w}}^{t+1} \right\|^2]
\end{aligned}
\quad (14)
$$

By the sampling scheme, one has

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}'_t}\left\|\tilde{\boldsymbol{w}}^{t+1}-\bar{\boldsymbol{w}}^{t+1}\right\|^2 &\leq \frac{1}{|\mathcal{S}'_t|}\mathbb{E}_i[\left\|\boldsymbol{w}_i^{t+1}-\bar{\boldsymbol{w}}^{t+1}\right\|^2] \\
&\leq \frac{1}{|\mathcal{S}'_t|}\mathbb{E}_i[\left\|\boldsymbol{w}_i^{t+1}-\boldsymbol{w}^t\right\|^2+\left\|\boldsymbol{w}^t-\bar{\boldsymbol{w}}^{t+1}\right\|^2+2(\boldsymbol{w}_i^{t+1}-\boldsymbol{w}^t)^\top(\boldsymbol{w}^t-\bar{\boldsymbol{w}}^{t+1})] \\
&\leq \frac{1}{|\mathcal{S}'_t|}\mathbb{E}_i[\left\|\boldsymbol{w}_i^{t+1}-\boldsymbol{w}^t\right\|^2 \qquad (\text{by } \mathbb{E}_i(\boldsymbol{w}_i^{t+1})=\bar{\boldsymbol{w}}^{t+1}) \\
&\overset{Eq.\ 8}{\leq} \frac{1}{|\mathcal{S}'_t|}\frac{(1+\gamma)^2}{\bar{\mu}^2}\mathbb{E}_i[\left\|\nabla F_i(\boldsymbol{w}^t)\right\|^2] \\
&\leq \frac{B^2}{|\mathcal{S}'_t|}\frac{(1+\gamma)^2}{\bar{\mu}^2}\left\|\nabla f(\boldsymbol{w}^t)\right\|^2 \quad (\text{by Assumption 4.1 (3)}). \qquad (15)
\end{aligned}$$

Combining Eq. 14, Eq. 15, and Eq. 9, together with the fact that $\mathbb{E}_{\mathcal{S}'_t}\left\|\tilde{\boldsymbol{w}}^{t+1}-\bar{\boldsymbol{w}}^{t+1}\right\| \leq \sqrt{\mathbb{E}_{\mathcal{S}'_t}\left\|\tilde{\boldsymbol{w}}^{t+1}-\bar{\boldsymbol{w}}^{t+1}\right\|^2}$ as a result of the Jesen's inequality, one reaches to

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}'_t}[f(\tilde{\boldsymbol{w}}^{t+1})] &\leq f(\bar{\boldsymbol{w}}^{t+1})+\left(2L\frac{B^2}{\sqrt{|\mathcal{S}'_t|}}\frac{(1+\gamma)^2}{\bar{\mu}^2}+\frac{B}{\sqrt{|\mathcal{S}'_t|}}\frac{(1+\gamma)}{\bar{\mu}}+L\frac{B^2}{|\mathcal{S}'_t|}\frac{(1+\gamma)^2}{\bar{\mu}^2}\right)\left\|\nabla f(\boldsymbol{w}^t)\right\|^2 \\
&= f(\bar{\boldsymbol{w}}^{t+1})+\left(\frac{LB^2(1+\gamma)^2}{|\mathcal{S}'_t|\bar{\mu}^2}(2\sqrt{|\mathcal{S}'_t|}+1)+\frac{B(1+\gamma)}{\bar{\mu}\sqrt{|\mathcal{S}'_t|}}\right)\left\|\nabla f(\boldsymbol{w}^t)\right\|^2. \qquad (16)
\end{aligned}$$

Combine Eq. 16 and Eq. 11, one reaches to

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}'_t}[f(\tilde{\boldsymbol{w}}^{t+1})] \leq f(\boldsymbol{w}^t)-\Bigg(\frac{1-\gamma B}{\mu}-\frac{LB(1+\gamma)}{\mu\bar{\mu}}-\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}- \\
\left(\frac{LB^2(1+\gamma)^2}{|\mathcal{S}'_t|\bar{\mu}^2}(2\sqrt{|\mathcal{S}'_t|}+1)+\frac{B(1+\gamma)}{\bar{\mu}\sqrt{|\mathcal{S}'_t|}}\right)\Bigg)\left\|\nabla f(\boldsymbol{w}^t)\right\|^2 \qquad (17)
\end{aligned}$$

Taking the expectation with respect to the discretization mechanism, $\mathbb{E}_{\mathcal{M}}[\boldsymbol{w}^{t+1}] = \tilde{\boldsymbol{w}}^{t+1}$ by Lemma 4.1. Since f is L-smooth,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{M}}[f(\boldsymbol{w}^{t+1})] &\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{L}{2}\mathbb{E}_{\mathcal{M}}[\left\|\boldsymbol{w}^{t+1} - \tilde{\boldsymbol{w}}^{t+1}\right\|^2] \\
&= f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{L}{2}\frac{1}{|\mathcal{S}_t'|^2}\mathbb{E}_{\mathcal{M}}\left[\left\|\sum_{i\in\mathcal{S}_t'}\left(\mathcal{M}(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}) - (\boldsymbol{w}_i^{t+1} - \boldsymbol{w})\right)\right\|^2\right] \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{L}{8|\mathcal{S}_t'|^2}\left\|\sum_{i\in\mathcal{S}_t'}(\boldsymbol{w}_i^{t+1} - \boldsymbol{w})\right\|^2 \qquad \text{(by Lemma 4.1)} \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{L}{8|\mathcal{S}_t'|^2}\left(\sum_{i\in[n]}\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|^2 + \sum_{i\neq j}(\boldsymbol{w}_i^{t+1} - \boldsymbol{w})^\top(\boldsymbol{w}_j^{t+1} - \boldsymbol{w})\right) \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{L}{8|\mathcal{S}_t'|^2}\left(\sum_{i\in[n]}\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|^2 + \sum_{i\neq j}\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|\left\|\boldsymbol{w}_j^{t+1} - \boldsymbol{w}\right\|\right) \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{LN}{8|\mathcal{S}_t'|^2}\left(\sum_{i\in[n]}\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|^2\right) \qquad \text{(by } 2\left\|a\right\|\left\|b\right\| \leq \left\|a\right\|^2 + \left\|b\right\|^2\text{)} \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{LN}{8|\mathcal{S}_t'|^2 p_{\min}}\left(\sum_{i\in[n]}p_{\min}\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|^2\right) \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{LN}{8|\mathcal{S}_t'|^2 p_{\min}}\left(\sum_{i\in[n]}p_i\left\|\boldsymbol{w}_i^{t+1} - \boldsymbol{w}\right\|^2\right) \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{LN}{8|\mathcal{S}_t'|^2 p_{\min}}\left(\sum_{i\in[n]}p_i\frac{(1+\gamma)^2}{\bar{\mu}^2}\left\|\nabla F_i(\boldsymbol{w}^t)\right\|^2\right) \\
&\leq f(\tilde{\boldsymbol{w}}^{t+1}) + \frac{LN}{8|\mathcal{S}_t'|^2 p_{\min}}\frac{B^2(1+\gamma)^2}{\bar{\mu}^2}\left\|\nabla f(\boldsymbol{w}^t)\right\|^2
\end{aligned} \tag{18}
$$

Put Eq. 18 and Eq. 17 together, we reach to

$$
\begin{aligned}
\mathbb{E}_{\mathcal{M},\mathcal{S}_t'}[f(\boldsymbol{w}^{t+1})] \leq f(\boldsymbol{w}^t) - \Bigg(&\frac{1-\gamma B}{\mu} - \frac{LB(1+\gamma)}{\mu\bar{\mu}} - \frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2} - \\
&\left(\frac{LB^2(1+\gamma)^2}{|\mathcal{S}_t'|\bar{\mu}^2}(2\sqrt{|\mathcal{S}_t'|}+1) + \frac{B(1+\gamma)}{\bar{\mu}\sqrt{|\mathcal{S}_t'|}}\right) - \frac{LNB^2(1+\gamma)^2}{8|\mathcal{S}_t'|^2 p_{\min}\bar{\mu}^2}\Bigg)\left\|\nabla f(\boldsymbol{w}^t)\right\|^2
\end{aligned} \tag{19}
$$

When there is no adversary, then $|\mathcal{S}'| = K$, then

$$
\mathbb{E}_{\mathcal{M},\mathcal{S}_t'}[f(\boldsymbol{w}^{t+1})] \leq f(\boldsymbol{w}^t) - \kappa\left\|\nabla f(\boldsymbol{w}^t)\right\|^2
$$

Finally, taking the total expectation with respect to all randomness and by telescoping, one reaches

$$
\kappa\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|]^2 \leq f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*).
$$

Divide T on both sides, then

$$
\min_{t\in[T-1]}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|]^2 \leq \frac{f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*)}{\kappa T}.
$$

□