Appendix

Proof of Eq. 2

Proof. Let Δ denote the local update before clipping and adding noise, $\widetilde{\Delta}_i^t = \Delta_i^t \cdot \min(1, \frac{S}{\|\Delta_i^t\|})$ denote the local update before clipping but after adding noise, $\overline{\Delta}_i^t = \widetilde{\Delta}_i^t + \mathcal{N}(0, \frac{\sigma^2 S^2}{|\mathcal{P}_t|})$ denote the local update after clipping and adding noise. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{d}\left\|\overline{\Delta}_i^t - \Delta_i^t\right\|_2^2\right] \\
&\leq \frac{1}{d}\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t - \Delta_i^t\right\|_2^2\right] + \frac{1}{d}\mathbb{E}\left[\left\|\overline{\Delta}_i^t - \widetilde{\Delta}_i^t\right\|_2^2\right] \\
&= \frac{1}{d}\mathbb{E}\left[\left\|\Delta_i^t \cdot \min(1, \frac{S}{\|\Delta_i^t\|}) - \Delta_i^t\right\|_2^2\right] + \frac{1}{d}\mathbb{E}\left[\left\|\mathcal{N}(0, \frac{\sigma^2 S^2}{|\mathcal{P}_t|}\mathbf{I}_d)\right\|_2^2\right] \\
&= \frac{1}{d}\mathbb{E}\left[\left\|\frac{\Delta_i^t}{\|\Delta_i^t\|} \cdot \min(0, S - \|\Delta_i^t\|)\right\|_2^2\right] + \frac{1}{d}\mathbb{E}\left[d \cdot \mathcal{N}(0, \frac{\sigma^2 S^2}{|\mathcal{P}_t|})^2\right] \\
&= \frac{1}{d}\min(0, S - \|\Delta_i^t\|)^2 + \mathbb{E}\left[\mathcal{N}(0, \frac{\sigma^2 S^2}{|\mathcal{P}_t|})^2\right] \\
&= \frac{1}{d}\max\left(0, \|\Delta_i^t\| - S\right)^2 + \frac{\sigma^2 S^2}{|\mathcal{P}_t|} \qquad (14)
\end{aligned}
$$

Proof of Lemma 2

Proof. According to Eq. 5 and Eq. 7, the local update at step Q of agent i at round t can be written as

$$
\begin{aligned}
\mathbf{w}_i^{t,Q} &= \mathbf{w}_i^{t,Q-1} - \eta_l\left(\mathbf{g}_i^{t,Q} + \lambda(\mathbf{w}_i^{t,Q-1} - \mathbf{w}^t)\right) \\
&= (1 - \lambda\eta_l)\mathbf{w}_i^{t,Q-1} - \eta_l \mathbf{g}_i^{t,Q} + \lambda\eta_l \mathbf{w}^t \qquad (15)
\end{aligned}
$$

where $\mathbf{g}_i^{t,q} = \frac{1}{|\mathcal{B}_q|}\sum_{(x,y)\in\mathcal{B}_q}\nabla f_i\left(\mathbf{w}_i^{t,q-1}, x, y\right)$. Unrolling Eq. 15, we have

$$
\begin{aligned}
\mathbf{w}_i^{t,Q} = &-\eta\sum_{q=0}^{Q-1}\left[(1-\lambda\eta_l)^q \mathbf{g}_i^{t,q}\right] \\
&+ \underbrace{\left[(1-\lambda\eta_l)^Q + \sum_{q=0}^{Q-1}(1-\lambda\eta_l)^q\lambda\eta_l\right]}_{A_1} \cdot \mathbf{w}^t \qquad (16)
\end{aligned}
$$

Simplifying A_1, we have

$$
\begin{aligned}
A_1 &= (1-\lambda\eta_l)^Q + \sum_{q=0}^{Q-1}(1-\lambda\eta_l)^q\lambda\eta_l \\
&= (1-\lambda\eta_l)^Q + (1-\lambda\eta_l)^{Q-1}\lambda\eta_l + \sum_{q=0}^{Q-2}(1-\lambda\eta_l)^q\lambda\eta_l \\
&= (1-\lambda\eta_l)^{Q-1}(1-\lambda\eta_l+\lambda\eta_l) + \sum_{q=0}^{Q-2}(1-\lambda\eta_l)^q\lambda\eta_l \\
&= (1-\lambda\eta_l)^{Q-1} + \sum_{q=0}^{Q-2}(1-\lambda\eta_l)^q\lambda\eta_l \\
&= (1-\lambda\eta_l)^{Q-2} + \sum_{q=0}^{Q-3}(1-\lambda\eta_l)^q\lambda\eta_l \\
&\cdots \\
&= (1-\lambda\eta_l)^1 + \sum_{q=0}^{0}(1-\lambda\eta_l)^q\lambda\eta_l \\
&= 1 \qquad (17)
\end{aligned}
$$

Replacing A_1 in Eq. 16 yields the desired result.

Proof of Theorem 1

Our Theorem 1 is directly obtained by applying [[2], Theorem 1] to the global learning process of our protocol, with sampling probability $q = \frac{P}{N}$ and global steps T. Refering to [2] for more details.

Proof of Theorem 2

Proof. For convenience, we first define the following notations:

$$
\begin{aligned}
\Delta_i^t &:= -\eta_l\sum_{q=0}^{Q-1}\gamma_i^{t,q}\mathbf{g}_i^{t,q}\cdot\alpha_i^t, \\
\overline{\Delta}_i^t &:= -\eta_l\sum_{q=0}^{Q-1}\gamma_i^{t,q}\mathbf{g}_i^{t,q}\cdot\overline{\alpha}^t, \\
\breve{\Delta}_i^t &:= -\eta_l\sum_{q=0}^{Q-1}\gamma_i^{t,q}\nabla f_i(x_i^{t,q})\cdot\overline{\alpha}^t
\end{aligned}
$$

where

$$
\alpha_i^t := \min\left(1, \frac{S}{\eta_l\beta_i^t\left\|\sum_{q=0}^{Q-1}\gamma_i^{t,q}\mathbf{g}_i^{t,q}\right\|}\right), \overline{\alpha}^t := \frac{1}{N}\sum_{i=1}^{N}\alpha_i^t,
$$

$$
\beta_i^t = \frac{\|M_i^t \circ \sum_{q=0}^{Q-1}\gamma_i^{t,q}\mathbf{g}_i^{t,q}\|}{\|\sum_{q=0}^{Q-1}\gamma_i^{t,q}\mathbf{g}_i^{t,q}\|}
$$

By Lipschitz smoothness, we have

$$
\begin{aligned}
&\mathbb{E}\left[f\left(x_{t+1}\right)\right] \\
&\leq f\left(x_t\right)+\left\langle \nabla f\left(x_t\right), x_{t+1}-x_t\right\rangle+\frac{L}{2}\left\|x_{t+1}-x_t\right\|^2 \\
&=f\left(x_t\right)+\eta_g\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t+z_i^t\right]\right\rangle \\
&+\frac{L}{2}\eta_g^2\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t+z_i^t\right\|^2\right] \\
&=f\left(x_t\right)+\eta_g\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t\right]\right\rangle \\
&+\frac{L}{2}\eta_g^2\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t\right\|^2\right]+\frac{\eta_g^2 L\sigma^2 S^2 d}{2P^2} \qquad (18)
\end{aligned}
$$

where d represents dimension of $x_i^{t,q}$; in the last equation we use the fact that z_i^t is zero mean. Next, we will analyse the bias caused by clipping, through analyzing the first order term in the above expression. Towards this end, we have the following series of relations:

$$
\begin{aligned}
&\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t\right]\right\rangle \\
&=\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{P}\mathbb{E}_i\left[\sum_{i\in\mathcal{P}_t}\Delta_i^t\right]\right]\right\rangle \\
&=\left\langle \nabla f\left(x_t\right), \frac{1}{P}P\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\Delta_i^t\right]\right\rangle \\
&=\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\Delta_i^t-\overline{\Delta}_i^t\right]\right\rangle \\
&+\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\overline{\Delta}_i^t\right]\right\rangle \qquad (19)
\end{aligned}
$$

Then we bound the two terms in the above expression, respectively. To bound the first term, we have:

$$
\begin{aligned}
&\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\Delta_i^t-\overline{\Delta}_i^t\right]\right\rangle \\
&=\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[-\frac{1}{N}\sum_{i=1}^N\sum_{q=0}^{Q-1}\eta_l(\alpha_i^t-\overline{\alpha}^t)\gamma_i^{t,q}\mathbf{g}_i^{t,q}\right]\right\rangle \\
&\leq\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\sum_{q=0}^{Q-1}\eta_l|\alpha_i^t-\overline{\alpha}^t|\gamma_i^{t,q}\mathbf{g}_i^{t,q}\right]\right\rangle \\
&\leq\left\langle \nabla f\left(x_t\right), \eta_l Q\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t|\mathbf{g}_i^{t,q}\right]\right\rangle \\
&=\frac{\eta_l Q}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t|\left\langle \nabla f\left(x_t\right), \mathbf{g}_i^{t,q}\right\rangle \\
&=\frac{\eta_l Q}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t|\left\langle \frac{1}{N}\sum_{j=1}^N\nabla f_j\left(x_t\right), \mathbf{g}_i^{t,q}\right\rangle \\
&=\frac{\eta_l Q}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t|\left(\frac{1}{N}\left\langle \nabla f_i\left(x_t\right), \mathbf{g}_i^{t,q}\right\rangle\right) \\
&\leq\frac{\eta_l Q G^2}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t| \\
&=\eta_l\tilde{\alpha}^t Q G^2 \qquad (20)
\end{aligned}
$$

where $\tilde{\alpha}^t=\frac{1}{N}\sum_{i=1}^N|\alpha_i^t-\overline{\alpha}^t|$. To bound the second term, we have:

$$
\begin{aligned}
&\left\langle \nabla f\left(x_t\right), \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\overline{\Delta}_i^t\right]\right\rangle=\mathbb{E}\left[\left\langle \nabla f\left(x_t\right), \frac{1}{N}\sum_{i=1}^N\breve{\Delta}_i^t\right\rangle\right] \\
&=\frac{-\eta_l\overline{\alpha}^t Q}{2}\left\|\nabla f\left(x_t\right)\right\|^2-\frac{\eta_l\overline{\alpha}^t}{2Q}\mathbb{E}\left[\left\|\frac{1}{\eta_l\overline{\alpha}^t N}\sum_{i=1}^N\breve{\Delta}_i^t\right\|^2\right] \\
&+\underbrace{\frac{\eta_l\overline{\alpha}^t}{2}\mathbb{E}\left[\left\|\sqrt{Q}\nabla f\left(x_t\right)-\frac{1}{\eta_l\overline{\alpha}^t N\sqrt{Q}}\sum_{i=1}^N\breve{\Delta}_i^t\right\|^2\right]}_{A_2} \qquad (21)
\end{aligned}
$$

where the second equation is because $\langle a,b\rangle=-\frac{1}{2}\|a\|^2-\frac{1}{2}\|b\|^2+\frac{1}{2}\|a-b\|^2$ holds true for any vector a and b. Next,

we bound A_2 as follows:

$$
\begin{aligned}
A &= Q\mathbb{E}\left[\left\|\nabla f\left(x_t\right)-\frac{1}{QN}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\gamma_i^{t,q}\nabla f_i\left(x_i^{t,q}\right)\right\|^2\right] \\
&= Q\mathbb{E}\left[\left\|\frac{1}{QN}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\nabla f_i\left(x^t\right)-\gamma_i^{t,q}\nabla f_i\left(x_i^{t,q}\right)\right\|^2\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\mathbb{E}\left[\left\|\nabla f_i\left(x^t\right)-\gamma_i^{t,q}\nabla f_i\left(x_i^{t,q}\right)\right\|^2\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\mathbb{E}\left[\left\|\nabla f_i\left(x^t\right)-\nabla f_i\left(x_i^{t,q}\right)\right\|^2\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\mathbb{E}\left[\left\|\nabla f_i\left(x_i^{t,q}\right)-\gamma_i^{t,q}\nabla f_i\left(x_i^{t,q}\right)\right\|^2\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{Q-1}\left(L^2\mathbb{E}\left[\left\|x^t-x_i^{t,q}\right\|^2\right]+\mathbb{E}\left[(1-(\gamma_i^{t,q}))^2G^2\right]\right) \\
&\leq L^2 5Q^2\eta_l^2\left(\sigma_l^2+6Q\sigma_g^2\right)+L^2 30Q^3\eta_l^2\left\|\nabla f\left(x_t\right)\right\|^2+QG^2
\end{aligned}
\tag{22}
$$

where the first inequality comes from Jensen's inequality, the third inequality comes from L-smoothness, and the last inequality is due to [[29], Lemma 3] which indicates that the following inequality holds for any q

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|x^t-x_i^{t,q}\right\|^2\right] \\
&\leq 5Q\eta_l^2\left(\sigma_l^2+6Q\sigma_g^2\right)+30Q^2\eta_l^2\left\|\nabla f\left(x_t\right)\right\|^2
\end{aligned}
\tag{23}
$$

Next, we turn to upper bounding the second order term in Eq. 18 as follows

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i\in\mathcal{P}_t}\Delta_i^t\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i\in\mathcal{P}_t}\sum_{q=0}^{Q-1}\eta_l\alpha_i^t\gamma_i^{t,q}\mathbf{g}_i^{t,q}\right\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{P}\sum_{i\in\mathcal{P}_t}\sum_{q=0}^{Q-1}(\eta_l\alpha_i^t\gamma_i^{t,q})^2\left\|\mathbf{g}_i^{t,q}\right\|^2\right] \\
&\leq \frac{\eta_l^2G^2}{P^2}\mathbb{E}\left[\sum_{i\in\mathcal{P}_t}\sum_{q=0}^{Q-1}(\alpha_i^t\gamma_i^{t,q})^2\right] \\
&\leq \frac{\eta_l^2G^2Q}{P}
\end{aligned}
\tag{24}
$$

where the first inequation is because the bounded gradient assumption.

Combining Eq. 18-24, we have

$$
\begin{aligned}
&\mathbb{E}\left[f\left(x_{t+1}\right)\right] \leq f\left(x_t\right)+\eta_g\eta_l\tilde{\alpha}^tQG^2 \\
&-\frac{\eta_g\eta_l\overline{\alpha}^tQ}{2}\left\|\nabla f\left(x_t\right)\right\|^2-\frac{\eta_g\eta_l\overline{\alpha}^t}{2Q}\mathbb{E}\left[\left\|\frac{1}{\eta_lN\overline{\alpha}^t}\sum_{i=1}^{N}\breve{\Delta}_i^t\right\|^2\right] \\
&+\frac{\eta_g\eta_l\overline{\alpha}^t}{2}\left(5L^2Q^2\eta_l^2\left(\sigma_l^2+6Q\sigma_g^2\right)+30L^2Q^3\eta_l^2\left\|\nabla f\left(x_t\right)\right\|^2+QG^2\right) \\
&+\frac{\eta_l^2\eta_g^2LG^2Q}{2P}+\frac{\eta_g^2L\sigma^2S^2d}{2P^2}
\end{aligned}
\tag{25}
$$

When $\eta_l \leq \frac{1}{\sqrt{60}QL}$ the above inequality simplifies to

$$
\begin{aligned}
&\mathbb{E}\left[f\left(x_{t+1}\right)\right] \leq f\left(x_t\right)-\frac{\eta_g\eta_l\overline{\alpha}^tQ}{4}\left\|\nabla f\left(x_t\right)\right\|^2+\eta_g\eta_l\tilde{\alpha}^tQG^2 \\
&+\frac{5\eta_g\eta_l^3\overline{\alpha}^tL^2Q^2}{2}\left(\sigma_l^2+6Q\sigma_g^2\right)+\frac{\eta_g\eta_l\overline{\alpha}^tQG^2}{2} \\
&+\frac{\eta_l^2\eta_g^2LG^2Q}{2P}+\frac{\eta_g^2L\sigma^2S^2d}{2P^2}
\end{aligned}
\tag{26}
$$

Sum over t from 1 to T, divide both side by $\frac{\eta_g\eta_lQT}{4}$, and rearrange, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(x_t\right)\right\|^2\right] \leq \frac{4}{\eta_g\eta_lQT}\mathbb{E}\left[f\left(x_1\right)-f\left(x_{T+1}\right)\right] \\
&+\frac{4G^2}{T}\sum_{t=1}^{T}\tilde{\alpha}^t+\frac{10\eta_l^2L^2Q}{T}\left(\sigma_l^2+6Q\sigma_g^2\right)\sum_{t=1}^{T}\overline{\alpha}^t+\frac{2G^2}{T}\sum_{t=1}^{T}\overline{\alpha}^t \\
&+\frac{2\eta_l\eta_gLG^2}{P}+\frac{2\eta_g L\sigma^2S^2d}{\eta_lQP^2}
\end{aligned}
\tag{27}
$$

As both $\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ and $\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t$ are bounded, the big-$\mathcal{O}$ convergence about T, Q, P, η_l, η_g is

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(x^t\right)\right\|^2\right] \\
&\leq \mathcal{O}\left(\frac{1}{\eta_g\eta_lQT}+\eta_l^2Q^2+\frac{\eta_g\eta_l}{P}+\frac{\eta_g\sigma^2S^2d}{\eta_lQP^2}\right)
\end{aligned}
\tag{28}
$$

which yield the result.