

## A THEORETICAL ANALYSIS

Our proof follows similar lines to that of Cho et al. (2022) but with modifications based on our problem formulation of having good and bad clients as well as our different skewness metrics and local-global objective gap  $\rho_g, \rho_b$ , and  $\Gamma_g$ , respectively. To begin, we present some preliminary lemmas that are useful for the proof of Theorem 1.

### A.1 PRELIMINARY LEMMAS

**Lemma 1.** Assume  $F_k$  is  $L$ -smooth with global optimum at  $w_k^*$ . Then for any  $w_k$  in the domain of  $F_k$ ,

$$\|\nabla F_k(w_k)\|^2 \leq 2L(F_k(w_k) - F_k(w_k^*)).$$

*Proof.* Since  $F_k$  is  $L$ -smooth,

$$F_k(w_k) - F_k(w_k^*) - \langle \nabla F_k(w_k^*), w_k - w_k^* \rangle \geq \frac{1}{2L} \|\nabla F_k(w_k) - \nabla F_k(w_k^*)\|^2$$

and  $\nabla F_k(w_k^*) = 0$  since  $w_k^*$  is a minimizer, so this implies

$$F_k(w_k) - F_k(w_k^*) \geq \frac{1}{2L} \|\nabla F_k(w_k)\|^2$$

which yields the claim.  $\square$

**Lemma 2.** For  $w_k^{(t)}$  and  $\bar{w}^{(t)} = \frac{1}{m} \sum_{k \in S^{(t)}} w_k^{(t)}$ ,

$$\frac{1}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 \right] \leq 16\eta_t^2 \tau^2 G^2.$$

*Proof.* We have

$$\begin{aligned} \frac{1}{m} \sum_{k \in S^{(t)}} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 &\leq \sum_{k \in S^{(t)}} \left\| \frac{1}{m} \sum_{\tilde{k} \in S^{(t)}} w_{\tilde{k}}^{(t)} - w_k^{(t)} \right\|^2 = \frac{1}{m^2} \sum_{k \in S^{(t)}} \sum_{\tilde{k} \in S^{(t)}} \|w_{\tilde{k}}^{(t)} - w_k^{(t)}\|^2 \\ &= \frac{1}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \|w_{\tilde{k}}^{(t)} - w_k^{(t)}\|^2, \end{aligned}$$

where the inequality follows from  $\|\sum_{i=1}^n \mathbf{x}_i\|^2 \leq n \sum_{i=1}^n \|\mathbf{x}_i\|^2$ . For  $k = \tilde{k}$ , the right hand side of the above inequality is zero. Since the selected clients get updated at every  $\tau$  for any  $t$  there exist  $t_0$  such that  $w_{\tilde{k}}^{(t_0)} = w_k^{(t)}$ , where  $0 \leq t - t_0 \leq \tau$ . Hence for any  $t$ ,  $\|w_{\tilde{k}}^{(t)} - w_k^{(t)}\|^2$  is bounded above by  $\tau$  updates. With non-increasing  $\eta_t$  over  $t$  and  $\eta_{t_0} \leq 2\eta_t$ , we can write the right hand side of the above inequality as

$$\begin{aligned} \frac{1}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \|w_{\tilde{k}}^{(t)} - w_k^{(t)}\|^2 &\leq \frac{1}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \left\| \sum_{i=t_0}^{t_0+\tau-1} \eta_i \left( g_{\tilde{k}}(w_{\tilde{k}}^{(i)}, \xi_{\tilde{k}}^{(i)}) - g_k^{(i)}(w_k^{(i)}, \xi_k^{(i)}) \right) \right\|^2 \\ &\leq \frac{\eta_{t_0}^2 \tau}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \sum_{i=t_0}^{t_0+\tau-1} [2\|g_{\tilde{k}}(w_{\tilde{k}}^{(i)}, \xi_{\tilde{k}}^{(i)})\|^2 + 2\|g_k^{(i)}(w_k^{(i)}, \xi_k^{(i)})\|^2]. \end{aligned}$$

Taking expectation and applying Assumption 4 gives

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \|w_k^{(t)} - w_{\tilde{k}}^{(t)}\|^2\right] &\leq \frac{2\eta_{t_0}^2 \tau}{m^2} \mathbb{E}\left[\sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \sum_{i=t_0}^{t_0+\tau-1} \left[\|g_{\tilde{k}}(w_k^{(i)}, \xi_k^{(i)})\|^2 + \|g_k^{(i)}(w_{\tilde{k}}^{(i)}, \xi_{\tilde{k}}^{(i)})\|^2\right]\right] \\
&\leq \frac{8\eta_t^2 \tau}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} \sum_{i=t_0}^{t_0+\tau-1} (G^2 + G^2) = \frac{8\eta_t^2 \tau}{m^2} \sum_{\substack{k, \tilde{k} \in S^{(t)} \\ k \neq \tilde{k}}} 2\tau G^2 \\
&= \frac{8\eta_t^2 \tau}{m^2} m(m-1)2\tau G^2 \leq 16\eta_t^2 \tau^2 G^2.
\end{aligned}$$

□

**Lemma 3.** For any random selection strategy,  $\mathbb{E}\|\bar{w}^{(t)} - w^*\|^2$  has the following upper bound:

$$\mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] \leq \frac{1}{m} \mathbb{E}\left[\sum_{k \in S^{(t)}} \|w_k^{(t)} - w^*\|^2\right].$$

*Proof.*

$$\begin{aligned}
\mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] &= \mathbb{E}\left[\left\|\frac{1}{m} \sum_{k \in S^{(t)}} w_k^{(t)} - w^*\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{m} \sum_{k \in S^{(t)}} (w_k^{(t)} - w^*)\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{m} \sum_{k \in S^{(t)}} \|w_k^{(t)} - w^*\|^2\right].
\end{aligned}$$

□

## A.2 PROOF OF THEOREM 1

Letting  $\bar{g}(t) = \frac{1}{m} \sum_{k \in S^{(t)}} g_k(w_k^{(t)}, \xi_k^{(t)})$ , and using the condensed notation  $\bar{g}_k = \bar{g}_k(\bar{w}_k^{(t)}, \xi_k^{(t)})$  for simplicity, we have

$$\begin{aligned}
\|\bar{w}^{(t+1)} - w^*\|^2 &= \|\bar{w}^{(t)} - \eta_t \bar{g}^{(t)} - w^*\|^2 \\
&= \|\bar{w}^{(t)} - \eta_t \bar{g}^{(t)} - w^* - \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)}) + \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)})\|^2 \\
&= \|\bar{w}^{(t)} - w^* - \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)})\|^2 + \eta_t^2 \left\|\frac{1}{m} \sum_{k \in S^{(t)}} (\nabla F_k(w_k^{(t)}) - \bar{g}_k^{(t)})\right\|^2 \\
&\quad + 2\eta_t \langle \bar{w}^{(t)} - w^* - \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)}), \frac{1}{m} \sum_{k \in S^{(t)}} (\nabla F_k(w_k^{(t)}) - \bar{g}_k^{(t)}) \rangle \\
&= \|\bar{w}^{(t)} - w^*\|^2 - \underbrace{2\eta_t \langle \bar{w}^{(t)} - w^*, \frac{1}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)}) \rangle}_{A_1} \\
&\quad + \underbrace{2\eta_t \langle \bar{w}^{(t)} - w^* - \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)}), \frac{1}{m} \sum_{k \in S^{(t)}} (\nabla F_k(w_k^{(t)}) - \bar{g}_k^{(t)}) \rangle}_{A_2} + \underbrace{\eta_t^2 \left\|\frac{1}{m} \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)})\right\|^2}_{A_3} \\
&\quad + \underbrace{\eta_t^2 \left\|\frac{1}{m} \sum_{k \in S^{(t)}} (\nabla F_k(w_k^{(t)}) - \bar{g}_k^{(t)})\right\|^2}_{A_4} \\
&= \|\bar{w}^{(t)} - w^*\|^2 + A_1 + A_2 + A_3 + A_4.
\end{aligned} \tag{8}$$

We first bound the quantity  $A_1$  of inequality (8) as follows:

$$\begin{aligned}
A_1 &= -\frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle \bar{w}^{(t)} - w^*, \nabla F_k(w_k^{(t)}) \rangle \\
&= -\frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle \bar{w}^{(t)} - w_k^{(t)}, \nabla F_k(w_k^{(t)}) \rangle - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle w_k^{(t)} - w^*, \nabla F_k(w_k^{(t)}) \rangle \\
&\leq \frac{\eta_t}{m} \sum_{k \in S^{(t)}} \left( \frac{1}{\eta_t} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 + \eta_t \|\nabla F_k(w_k^{(t)})\|^2 \right) - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle w_k^{(t)} - w^*, \nabla F_k(w_k^{(t)}) \rangle \\
&\quad \text{(using the AM-GM and Cauchy-Schwarz inequalities)} \\
&= \frac{1}{m} \sum_{k \in S^{(t)}} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 + \frac{\eta_t^2}{m} \sum_{k \in S^{(t)}} \|\nabla F_k(w_k^{(t)})\|^2 - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle w_k^{(t)} - w^*, \nabla F_k(w_k^{(t)}) \rangle \\
&\leq \frac{1}{m} \sum_{k \in S^{(t)}} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 + \frac{2L\eta_t^2}{m} \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \\
&\quad - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \langle w_k^{(t)} - w^*, \nabla F_k(w_k^{(t)}) \rangle \quad \text{(using Lemma 1)} \\
&\leq \frac{1}{m} \sum_{k \in S^{(t)}} \|\bar{w}^{(t)} - w_k^{(t)}\|^2 + \frac{2L\eta_t^2}{m} \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \\
&\quad - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} \left[ F_k(w_k^{(t)}) - F_k(w^*) + \frac{\mu}{2} \|w_k^{(t)} - w^*\|^2 \right],
\end{aligned}$$

where the last inequality follows from  $\mu$  strong convexity of  $F_k$  (Assumption 2). Hence, by Lemma 2, the expected value of  $A_1$  satisfies

$$\begin{aligned}
\mathbb{E}[A_1] &\leq 16\eta_t^2 \tau^2 G^2 - \frac{\eta_t \mu}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} \|w_k^{(t)} - w^*\|^2 \right] + \frac{2L\eta_t^2}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \right] \\
&\quad - \frac{2\eta_t}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k(w^*)) \right]. \tag{9}
\end{aligned}$$

Leaving this bound aside for the moment, next notice that  $\mathbb{E}[A_2] = 0$  because of the unbiased gradient assumption (Assumption 3). We may then bound  $A_3$  by Lemma 2 as follows:

$$\begin{aligned}
\mathbb{E}[A_3] &= \mathbb{E} \left[ \frac{\eta_t^2}{m^2} \left\| \sum_{k \in S^{(t)}} \nabla F_k(w_k^{(t)}) \right\|^2 \right] \leq \frac{\eta_t^2}{m} \sum_{k \in S^{(t)}} \mathbb{E} \left[ \|\nabla F_k(w_k^{(t)})\|^2 \right] \\
&\leq \frac{2L\eta_t^2}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \right]. \tag{10}
\end{aligned}$$

Finally, the bound for  $A_4$  is as follows:

$$\begin{aligned}
\mathbb{E}[A_4] &= \mathbb{E} \left[ \frac{\eta_t^2}{m^2} \left\| \sum_{k \in S^{(t)}} (\nabla F_k(w_k^{(t)}) - g_k^{(t)}) \right\|^2 \right] = \frac{\eta_t^2}{m^2} \mathbb{E}_{S^{(t)}} \left[ \sum_{k \in S^{(t)}} \mathbb{E} \|\nabla F_k(w_k^{(t)}) - g_k^{(t)}\|^2 \right] \\
&\leq \frac{\eta_t^2 m \sigma^2}{m^2} = \frac{\eta_t^2 \sigma^2}{m}, \tag{11}
\end{aligned}$$

where the second equality and inequality use Assumption 3.

Using the bounds (9), (10), and (11) in (8), we have

$$\mathbb{E}[\|\bar{w}^{(t+1)} - w^*\|^2] \leq \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] + \sum_{i=1}^4 \mathbb{E}[A_i] \leq \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] - \frac{\eta_t \mu}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} \|w_k^{(t)} - w^*\|^2 \right]$$

$$\begin{aligned}
& + 16\eta_t^2 \tau^2 G^2 + \frac{\eta_t^2 \sigma^2}{m} + \frac{4L\eta_t}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \right] - \frac{2\eta_t}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k(w^*)) \right] \\
& \leq (1 - \eta_t \mu) \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] + 16\eta_t^2 \tau^2 G^2 + \frac{\eta_t^2 \sigma^2}{m} + \underbrace{\frac{4L\eta_t^2}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \right]}_{A_5} \\
& \underbrace{- \frac{2\eta_t}{m} \mathbb{E} \left[ \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k(w^*)) \right]}_{A_5}. \tag{12}
\end{aligned}$$

The final inequality above utilizes Lemma 3.

Now we bound  $A_5$  as follows:

$$\begin{aligned}
A_5 &= \mathbb{E} \left[ \frac{4L\eta_t^2}{m} \sum_{k \in S^{(t)}} F_k(w_k^{(t)}) - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} F_k(w_k^{(t)}) - \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} (F_k^* - F_k(w^*)) \right. \\
& \quad \left. + \frac{2\eta_t}{m} \sum_{k \in S^{(t)}} F_k^* - \frac{4L\eta_t^2}{m} \sum_{k \in S^{(t)}} F_k^* \right] \\
&= \mathbb{E} \left[ \underbrace{\frac{2\eta_t(2L\eta_t - 1)}{m} \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*)}_{A_6} + 2\eta_t \mathbb{E} \left[ \frac{1}{m} \sum_{k \in S^{(t)}} (F_k(w^*) - F_k^*) \right] \right].
\end{aligned}$$

Take  $\eta_t < 1/(4L)$  and define  $v_t = 2\eta_t(1 - 2L\eta_t) \geq 0$ ; then we can bound  $A_6$  as

$$\begin{aligned}
& - \frac{v_t}{m} \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k^*) \\
&= - \frac{v_t}{m} \sum_{k \in S^{(t)}} (F_k(w_k^{(t)}) - F_k(\bar{w}^{(t)}) + F_k(\bar{w}^{(t)}) - F_k^*) \\
&= - \frac{v_t}{m} \sum_{k \in S^{(t)}} [F_k(w_k^{(t)}) - F_k(\bar{w}^{(t)})] - \frac{v_t}{m} \sum_{k \in S^{(t)}} [F_k(\bar{w}^{(t)}) - F_k^*] \\
&\leq - \frac{v_t}{m} \sum_{k \in S^{(t)}} [\langle \nabla F_k(w_k^{(t)}), w_k^{(t)} - \bar{w}^{(t)} \rangle + \frac{\mu}{2} \|w_k^{(t)} - \bar{w}^{(t)}\|^2] - \frac{v_t}{m} \sum_{k \in S^{(t)}} [F_k(\bar{w}^{(t)}) - F_k^*] \\
&\leq \frac{v_t}{m} \sum_{k \in S^{(t)}} [\eta_t L (F_k(\bar{w}^{(t)}) - F_k^*) + (\frac{1}{2\eta_t} - \frac{\mu}{2}) \|w_k^{(t)} - \bar{w}^{(t)}\|^2] - \frac{v_t}{m} \sum_{k \in S^{(t)}} [F_k(\bar{w}^{(t)}) - F_k^*] \\
& \text{(using the Cauchy-Schwarz inequality, the AM-GM inequality, and Lemma 1)} \\
&= - \frac{v_t}{m} (1 - \eta_t L) \sum_{k \in S^{(t)}} (F_k(\bar{w}^{(t)}) - F_k^*) + \left( \frac{v_t}{2\eta_t m} - \frac{v_t \mu}{2m} \right) \sum_{k \in S^{(t)}} \|w_k^{(t)} - \bar{w}^{(t)}\|^2 \\
&\leq - \frac{v_t}{m} (1 - \eta_t L) \sum_{k \in S^{(t)}} (F_k(\bar{w}^{(t)}) - F_k^*) + \frac{1}{m} \sum_{k \in S^{(t)}} \|w_k^{(t)} - \bar{w}^{(t)}\|^2. \tag{13}
\end{aligned}$$

The first inequality above uses  $\mu$  strong convexity of  $F_k$ , the subsequent inequality uses  $L$ -smoothness of  $F_k$ , and the final inequality follows because  $\frac{v_t(1 - \eta_t \mu)}{2\eta_t} \leq 1$ . Hence, we can bound  $A_5$  as follows:

$$\begin{aligned}
\mathbb{E}[A_5] &\leq -\frac{\nu_t}{m}(1-\eta_t L)\mathbb{E}\left[\sum_{k \in S^{(t)}} (F_k(\bar{w}^{(t)}) - F_k^*)\right] + \frac{1}{m}\mathbb{E}\left[\sum_{k \in S^{(t)}} \|w_k^{(t)} - \bar{w}^{(t)}\|^2\right] \\
&\quad + \frac{2\eta_t}{m}\mathbb{E}\left[\sum_{k \in S^{(t)}} (F_k(w^*) - F_k^*)\right] \\
&\leq -\frac{\nu_t}{m}(1-\eta_t L)\mathbb{E}\left[\sum_{k \in S^{(t)}} (F_k(\bar{w}^{(t)}) - F_k^*)\right] + 16\eta_t^2 \tau^2 G^2 + \frac{2\eta_t}{m}\mathbb{E}\left[\sum_{k \in S^{(t)}} (F_k(w^*) - F_k^*)\right] \\
&= -\frac{\nu_t}{m}(1-\eta_t L)\mathbb{E}\left[\sum_{k \in S^{(t)} \cap \mathcal{G}} (F_k(\bar{w}^{(t)}) - F_k^*) + \sum_{k \in S^{(t)} \cap \mathcal{B}} (F_k(\bar{w}^{(t)}) - F_k^*)\right] + 16\eta_t^2 \tau^2 G^2 \\
&\quad + \frac{2\eta_t}{m}\mathbb{E}\left[\sum_{k \in S^{(t)} \cap \mathcal{G}} (F_k(w^*) - F_k^*) + \sum_{k \in S^{(t)} \cap \mathcal{B}} (F_k(w^*) - F_k^*)\right] \\
&= 16\eta_t^2 \tau^2 G^2 - \frac{\nu_t(1-\eta_t L)}{m}\mathbb{E}[(p\rho_g(S(\pi, \bar{w}^{\lceil t/\tau \rceil}), \bar{w}^{(t)})) \\
&\quad + q\rho_b(S(\pi, \bar{w}^{\lceil t/\tau \rceil}), \bar{w}^{(t)}))(F_g(\bar{w}^{(t)}) - \sum_{k \in \mathcal{G}} p_k F_k^*) + \frac{2\eta_t}{m}\mathbb{E}[(p\rho_g(S(\pi, \bar{w}^{\lceil t/\tau \rceil}), w^*) \\
&\quad + q\rho_b(S(\pi, \bar{w}^{\lceil t/\tau \rceil}), w^*))(F_g(w^*) - \sum_{k \in \mathcal{G}} p_k F_k^*)] \\
&\leq 16\eta_t^2 \tau^2 G^2 - \underbrace{\frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] (\mathbb{E}[F_g(\bar{w}^{(t)}) - \sum_{k \in \mathcal{G}} p_k F_k^*) + \frac{2\eta_t}{m} (p\tilde{\rho}_g + q\tilde{\rho}_b)\Gamma_g]}_{A_7} \tag{14}
\end{aligned}$$

We used the definition of  $\rho(S(\pi, w), w')$  and  $\Gamma_g$  to arrive at (14). We can get a bound for  $A_7$  in (14) as follows:

$$\begin{aligned}
A_7 &= -\frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \sum_{k \in \mathcal{G}} p_k (\mathbb{E}[F_k(\bar{w}^{(t)})] - F^* + F^* - F_k^*) \\
&= -\frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \sum_{k \in \mathcal{G}} p_k (\mathbb{E}[F_k(\bar{w}^{(t)})] - F^* + F^* - F_k^*) \\
&= -\frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \sum_{k \in \mathcal{G}} p_k (\mathbb{E}[F_k(\bar{w}^{(t)})] - F^*) \\
&\quad - \frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \sum_{k \in \mathcal{G}} p_k (F^* - F_k^*) \\
&= -\frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] (\mathbb{E}[F_g(\bar{w}^{(t)})] - F^*) - \frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \Gamma_g \\
&\text{(using the definition of } \Gamma_g) \\
&\leq -\frac{\nu_t(1-\eta_t L)\mu [p\bar{\rho}_g + q\bar{\rho}_b]}{2m} \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] - \frac{\nu_t(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \Gamma_g \\
&\text{(using } \mu \text{ strongly convexity)} \\
&= -\frac{2\eta_t(1-2L\eta_t)(1-\eta_t L)\mu [p\bar{\rho}_g + q\bar{\rho}_b]}{2m} \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] \\
&\quad - \frac{2\eta_t(1-2L\eta_t)(1-\eta_t L)}{m} [p\bar{\rho}_g + q\bar{\rho}_b] \Gamma_g \\
&\leq -\frac{3\eta_t\mu [p\bar{\rho}_g + q\bar{\rho}_b]}{8m} \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] - \frac{2\eta_t [p\bar{\rho}_g + q\bar{\rho}_b] \Gamma_g}{m} + \frac{6\eta_t^2 [p\bar{\rho}_g + q\bar{\rho}_b] L \Gamma_g}{m} \tag{15}
\end{aligned}$$

where equation (15) is due to  $\mu$  strong convexity and we used  $-2\eta_t(1 - 2L\eta_t)(1 - L\eta_t) \leq -\frac{3}{4}\eta_t$  and  $-(1 - 2L\eta_t)(1 - L\eta_t) \leq -(1 - 3L\eta_t)$ . Hence, the bound of  $A_5$  is as follows:

$$\begin{aligned} & \frac{4L\eta_t}{m} \mathbb{E} \left[ \sum_{k \in S(t)} [(F_k(w_k^{(t)}) - F_k^*) - \frac{2\eta_t}{m}(F_k(w_k^{(t)}) - F_k(w^*)))] \right] \\ & \leq -\frac{3\eta_t\mu[p\bar{\rho}_g + q\bar{\rho}_b]}{8m} \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] + \eta_t^2 \left( \frac{6[p\bar{\rho}_g + q\bar{\rho}_b]L\Gamma_g}{m} + 16\tau^2G^2 \right) \\ & \quad - \frac{2\eta_t[p\bar{\rho}_g + q\bar{\rho}_b]\Gamma_g}{m} + \frac{2\eta_t[p\tilde{\rho}_g + q\tilde{\rho}_b]\Gamma_g}{m}. \end{aligned} \quad (16)$$

Finally, using equation (12), and (16) we can bound  $\|\bar{w}^{(t+1)} - w^*\|$  as follows:

$$\begin{aligned} \mathbb{E}[\|\bar{w}^{(t+1)} - w^*\|] & \leq [1 - \eta_t\mu[1 + \frac{3(p\bar{\rho}_g + q\bar{\rho}_b)}{8m}]] \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] \\ & \quad + \eta_t^2 [32\tau^2G^2 + \frac{\sigma^2}{m} + \frac{6(p\bar{\rho}_g + q\bar{\rho}_b)L\Gamma_g}{m}] + \frac{2\eta_t\Gamma_g}{m}(p\tilde{\rho}_g + q\tilde{\rho}_b - p\bar{\rho}_g - q\bar{\rho}_b) \\ & \leq [1 - \eta_t\mu[1 + \frac{3(p\bar{\rho}_g + q\bar{\rho}_g)}{8m}]] \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] + \eta_t^2 [32\tau^2G^2 + \frac{\sigma^2}{m} + \frac{6(p\bar{\rho}_b + q\bar{\rho}_b)L\Gamma_g}{m}] \\ & \quad + \frac{2\eta_t\Gamma_g}{m}(p\tilde{\rho}_g + q\tilde{\rho}_b - p\bar{\rho}_g - q\bar{\rho}_g). \end{aligned} \quad (17)$$

Equation (17) is obtained using  $\bar{\rho}_g \leq \bar{\rho}_b$ , which gives

$$\begin{aligned} \mathbb{E}[\|\bar{w}^{(t+1)} - w^*\|] & \leq [1 - \eta_t\mu[1 + \frac{3\bar{\rho}_g}{8}]] \mathbb{E}[\|\bar{w}^{(t)} - w^*\|^2] \\ & \quad + \eta_t^2 [32\tau^2G^2 + \frac{\sigma^2}{m} + 6\bar{\rho}_bL\Gamma_g] + \frac{2\eta_t\Gamma_g}{m}(p\tilde{\rho}_g + q\tilde{\rho}_b - m\bar{\rho}_g). \end{aligned}$$

By setting  $\Delta_{t+1} = \mathbb{E}[\|\bar{w}^{(t+1)} - w^*\|^2]$ ,  $B = 1 + \frac{3\bar{\rho}_g}{8}$ ,  $C = 32\tau^2G^2 + \frac{\sigma^2}{m} + 6\bar{\rho}_bL\Gamma_g$ ,  $D = \frac{2\Gamma_g}{m}(p\tilde{\rho}_g + q\tilde{\rho}_b - m\bar{\rho}_g)$ , we get

$$\Delta_{t+1} \leq (1 - \eta_t\mu B)\Delta_t + \eta_t^2 C + D\eta_t.$$

For a decreasing stepsize,  $\eta_t = \frac{\beta}{t+\gamma}$  for some  $\beta > \frac{1}{\mu B}$ ,  $\gamma > 0$ , we have that  $\Delta_t \leq \frac{\psi}{t+\gamma}$ , where

$$\psi = \max \left\{ (\gamma + 1)\|\bar{w}^{(1)} - w^*\|^2, \frac{\beta^2 C + \beta D(t + \gamma)}{\beta\mu B - 1} \right\}.$$

This can be shown by induction on  $t$  (see Lemma 4 below). Then using the  $L$ -smoothness of  $F(\cdot)$  we get

$$\mathbb{E}[F(\bar{w}^{(t)})] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2} \frac{\psi}{\gamma + t}.$$

Now for  $\beta = \frac{1}{\mu}$ , we get

$$\begin{aligned} \mathbb{E}[F(\bar{w}^{(T)})] - F^* & \leq \frac{1}{(T + \gamma)} \left[ \frac{4L(32\tau^2G^2 + \sigma^2/m)}{3\mu^2\bar{\rho}_g} + \frac{8L^2\Gamma_g}{\mu^2} \frac{\bar{\rho}_b}{\bar{\rho}_g} + \frac{L(\gamma + 1)(\|\bar{w}^{(1)} - w^*\|^2)}{2} \right] \\ & \quad + \frac{8L\Gamma_g}{3\mu} \left( \frac{p\tilde{\rho}_g + q\tilde{\rho}_b}{m\bar{\rho}_g} - 1 \right), \end{aligned}$$

which completes the proof of the theorem.  $\square$

**Lemma 4.** For a decreasing stepsize,  $\eta_t = \frac{\beta}{t+\gamma}$  for some  $\beta > \frac{1}{\mu B}$ ,  $\gamma > 0$ ,

$$\Delta_t \leq \frac{\psi}{t+\gamma} \quad (18)$$

where,

$$\psi = \max\left\{(\gamma+1)\|\bar{w}^{(1)} - w^*\|^2, \frac{1}{\beta\mu B - 1}(\beta^2 C + D\beta(t+\gamma))\right\} \quad (19)$$

and

$$\Delta_{t+1} \leq (1 - \eta_t \mu B) \Delta_t + \eta_t^2 C + \eta_t D.$$

*Proof.* For  $t = 1$ , equation (18) holds clearly as (using (19))

$$\Delta_1 \leq \frac{\psi}{\gamma+1} \leq \|\bar{w}^{(1)} - w^*\|^2 = \Delta_1$$

Assume that it holds for some  $t$ , then

$$\begin{aligned} \Delta_{t+1} &\leq (1 - \eta_t \mu B) \Delta_t + \eta_t^2 C + \eta_t D \\ &\leq \left(1 - \frac{\beta}{t+\gamma} \mu B\right) \frac{\psi}{t+\gamma} + \frac{\beta^2}{(t+\gamma)^2} C + \frac{\beta}{t+\gamma} D \\ &= \frac{t+\gamma - \beta\mu B}{(t+\gamma)^2} \psi + \frac{\beta^2 C + \beta D(t+\gamma)}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2} \psi + \frac{\beta^2 C + \beta D(t+\gamma)}{(t+\gamma)^2} - \frac{\beta\mu B - 1}{(t+\gamma)^2} \psi \\ &= \frac{t+\gamma-1}{(t+\gamma)^2} \psi \quad (\text{Using (19)}) \\ &\leq \frac{t+\gamma-1}{(t+\gamma)^2 - 1} \psi = \frac{\psi}{t+\gamma+1} \end{aligned}$$

□

## B DATASET AND MODEL DESCRIPTION - EXTENDED

### B.1 DATASETS

We utilize four prominent datasets: MNIST, CIFAR10, FEMNIST, and the SHAKESPEARE dataset, widely referenced in the literature McMahan et al. (2017); Li et al. (2020c).

**MNIST LeCun et al. (2010).** Renowned for handwriting recognition, this dataset consists of 70,000 gray-scale  $28 \times 28$  images. It includes 60,000 training samples and 10,000 test samples, spanning ten classes (digits 0-9). We distribute MNIST training data evenly among 100 clients for the IID case. For Non-IID, each client possesses one dominant class with 80% of the data, while the remaining classes share 20%. In the extreme Non-IID scenario, a class contributes data to at most two clients. The standard test set evaluates global model performance.

**CIFAR10 Krizhevsky (2009).** Comprising 60,000 color  $32 \times 32$  images, the CIFAR10 dataset encompasses 50,000 training images and 10,000 test images across ten classes. Similar to MNIST, we consider three distribution types: IID, Non-IID, and extreme Non-IID. Dividing the dataset into 100 clients, each IID client receives 500 samples. For Non-IID scenarios, one dominant class constitutes 80% of a client's data, while the rest is shared among other classes. In the extreme Non-IID case, each class contributes data to a maximum of two clients. The test set is used to evaluate the performance of the global model

**FEMNIST Caldas et al. (2018).** Derived from the LEAF dataset and implemented using TensorFlow Federated, FEMNIST involves 3,383 unique users (first 1000 used). It offers 341,873 training

examples and 40,832 test examples, featuring gray-scale  $28 \times 28$  images. The dataset creates a non-IID and heterogeneous setting, with each user representing a distinct client. Test sets from distinct clients are collected together to evaluate global performance.

**SHAKESPEARE** Caldas et al. (2018). Based on "*The Complete Works of William Shakespeare*", this dataset uses speaking roles in plays to represent individual clients. It encompasses 715 genuine users (71 clients with at least 60 test data points), providing 16,068 training examples and 2,356 test examples in text format. Like FEMNIST, the SHAKESPEARE dataset is non-IID and heterogeneous, associating each user with a unique client.

Table 1: Dataset and Model

Dataset	Training	Test	#Client	Distribution	Model
MNIST	60,000	10,000	300	IID/Non-IID	LR
CIFAR10	50,000	10,000	300	IID/Non-IID	CNN
FEMNIST	341,873	40,832	3383	Non-IID	LR
SHAKESPEARE	16,068	2,356	715	Non-IID	RNN

## B.2 MODEL PARAMETERS

In the context of an edge setup with IoT devices as clients, we prioritize lightweight models to accommodate limited power and computational capabilities.

**MNIST.** For the MNIST dataset, we adopt a simple Multi-Layer Perceptron (MLP) classifier using TensorFlow Keras. The architecture includes two hidden layers with ReLU activation: one with 200 neurons, the other with 100 neurons. An output layer with 10 neurons and softmax activation handles classification. Input features are flattened, and labels are one-hot encoded. The model employs the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss. Training spans 300 epochs for all distribution cases.

**CIFAR10.** Employing the CIFAR10 dataset, we employ a lightweight Convolutional Neural Network (CNN) classifier using TensorFlow Keras. The CNN architecture involves two sets of convolutional layers, max-pooling layers, dropout layers, and fully connected layers. ReLU activation functions operate in the convolutional layers, while softmax is used for the output layer. The model employs categorical cross-entropy loss, the Adam optimizer with varying learning rates, and trains until 300 rounds.

**FEMNIST.** Addressing the FEMNIST dataset, we use a simple MLP with two hidden layers. These layers consist of fully connected dense layers with ReLU activation. The model input shape is 784 (pixels in each image), featuring 64 neurons in the first hidden layer. The output layer, with 10 neurons, lacks an activation function to complement the Sparse-Categorical-Crossentropy loss function. The optimization employs a learning rate of 0.001. Training spans 300 epochs.

**SHAKESPEARE.** Utilizing the SHAKESPEARE dataset, we deploy a Recurrent Neural Network (RNN) featuring a GRU layer with `stateful=True`. Input data is preprocessed via an ASCII character-to-index lookup table, forming sequences of length 50+1. The architecture integrates an embedding layer, a GRU layer with varying units, and a dense layer with 86 output units. A custom evaluation metric gauges the model’s character prediction accuracy across the input sequence. There are 150 rounds of training across the clients.

**Evaluation Scenarios** We consider three different scenarios that reflect potential data corruption due to sensor quality and aging:

**Label shuffling.** In this scenario, we consider sensors’ label interpretations are incorrect, leading to the assignment of random labels to data. We experiment with varying percentages of clients whose labels are randomly shuffled.

**Label flipping.** Here, a random label is assigned to each client, with the same labels across its data (e.g., all of Client 1’s data is labeled 2). We consider a fraction of sensors that consistently produce a fixed, random label output.

**Noisy data.** This scenario involves correct label interpretation but noisy feature spaces. To simulate this, we introduce Gaussian noise to the features. For selected clients, the input data is first normalized to  $[0, 1]$  and then we add Gaussian noise  $x = x + \epsilon$ , where  $\epsilon \sim N(0, 0.7)$ . The resulting values are clipped again to  $[0, 1]$ .



## C EVALUATION - EXTENDED

### C.1 COMPARISON WITH BENCHMARK ALGORITHMS - EXTENDED

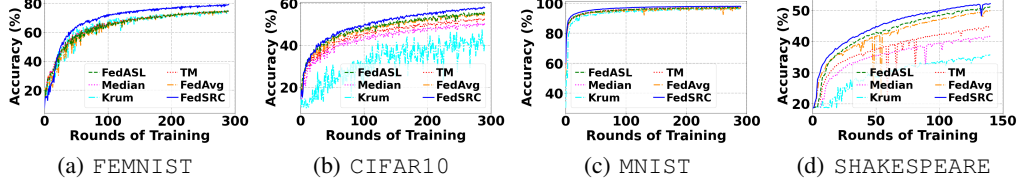


Figure 6: Comparison of global accuracy of FedSRC with other state of the arts algorithm for the FEMNIST, CIFAR10, MNIST and SHAKESPEARE datasets.

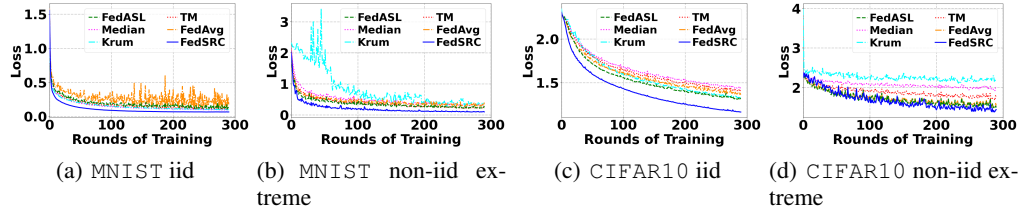


Figure 7: Comparison of loss of FedSRC with other state-of-the-arts algorithm for the CIFAR10 and MNIST datasets.

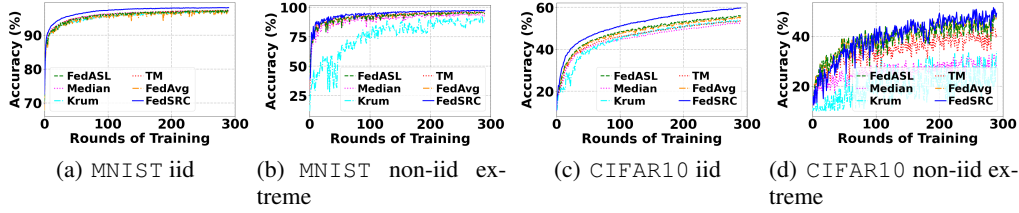


Figure 8: Comparison of accuracy of FedSRC with other state-of-the-arts algorithm for the CIFAR10 and MNIST datasets.

To comprehensively evaluate the effectiveness of FedSRC compared to state-of-the-art algorithms, we have conducted extensive assessments across diverse datasets, including FEMNIST, CIFAR10, MNIST, and SHAKESPEARE. These evaluations were carried out under our default settings, involving 30% data corruption. In this context, we present an in-depth evaluation focusing on MNIST and CIFAR10 datasets, considering both the IID (Independent and Identically Distributed) and Non-IID extreme cases.

In our evaluation, we blocked 30% of clients in FedSRC, Trimmed Mean, and Krum algorithms. In contrast, FedASL excludes clients falling outside one standard deviation, which accounts for discarding approximately 32% of clients. Notably, FedAVG does not discard any clients. The performance metrics displayed are the loss and accuracy of the global model when assessed against the test dataset.

Specifically, we present the outcomes in Figs. 7 represent the loss plot and 8 and 6 represent the accuracy plot. Our experiments reveal that FedSRC consistently outperforms other benchmark algorithms resulting in better global performance in the presence of corrupted clients.

### C.2 INTEGRATION WITH EXISTING ALGORITHMS

We demonstrate the effectiveness of integrating FedSRC with other algorithms by implementing it at the client level while maintaining aggregation protocols such as FedAVG, FedASL, Trimmed Mean,

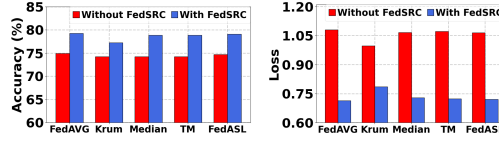


Figure 9: Performance comparison with integrating FedSRC with existing algorithms for FEMNIST.

Krum, and Median on the server side. As shown in Fig. 9, our integration approach enhances the performance of these pre-existing algorithms (about 6% increase in accuracy) and reduces the error loss (about 33% decrease in loss) in the presence of unreliable clients all the while also reducing computation and communication costs.

### C.3 SENSITIVITY ANALYSIS

**Impact of blocking percentage.** To understand the effects of user-defined blocking percentage, we evaluate the FEMNIST dataset with 30% data corruption. We vary client blocking from 0% to 90%. The goal is to find how the performance FedSRC is impacted by different degrees of client exclusion. The results, as depicted in Fig. 10, show that the optimal performance achieved when correctly estimating a 30% threshold. However, there is no significant degradation, especially when overestimating the blocking percentage.

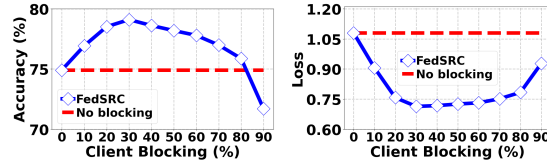


Figure 10: Effect of our cutoff (range) in performance of FedSRC for FEMNIST dataset.

The results highlight the robustness of our approach.

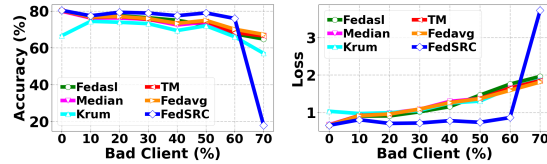


Figure 11: Performance comparison of FedSRC with other algorithms in the presence of different percentages of bad clients for FEMNIST dataset in shuffling.

**Impact of Different Percentage of Bad Client:** To assess our algorithm against varying levels of corrupted data, we use FEMNIST dataset with different percentages of bad clients and set the client blocking parameters of FedSRC and benchmark algorithms. Fig. 11 shows that as the percentage of unreliable clients increases, conventional algorithms' accuracy declines. In contrast, our FedSRC demonstrates remarkable robustness, effectively managing up to 60% of clients with erroneous behavior. Naturally, as our algorithm utilizes clients' loss statistics, its performance falters drastically with a higher percentage of bad clients.