

A PROOFS

Theorem 1 (Impossibility Theorem). *Let \mathcal{H} be a non-trivial hypothesis space and $\mathcal{L} : (\mathcal{X}, \mathcal{Y})^{(m \times n)} \rightarrow \mathcal{H}$ be the learner for an FL system. There exists a client participation process \mathcal{F} , a distribution P , and a target function $f \in \mathcal{H}$ with $\min_{h \in \mathcal{H}} \mathcal{R}_P(h, f) = 0$, such that $\mathbb{P}_{S \sim P}[\mathcal{R}_P(\mathcal{L}(\mathcal{F}(S)), f)] > \frac{1-\alpha}{8}] > \frac{1}{20}$.*

Proof. Denote S the dataset with size Mn i.i.d. sampled from distribution P , $\mathcal{F}(\cdot)$ the sampling process of FL system, and $\bar{S} = \mathcal{F}(S)$ the training dataset selected by FL system with size mn . Consider a distribution P with support on only two points $\{x_1, x_2\}$ such that $\mathbb{P}_P(x_1) = 1 - 4\epsilon$ and $\mathbb{P}_P(x_2) = 4\epsilon$ with $\epsilon = \frac{1-\alpha}{8}$.

First we show that the rare points x_2 appears at most $(1 - \alpha)Mn$ times with constant probability. Let \hat{s} be the number of x_2 points in S , then $\hat{s} \sim \mathbb{B}(Mn, \epsilon)$ is a binomial random variable. By the Chernoff bound,

$$\mathbb{P}[\hat{s} \geq (1 - \alpha)Mn] = \mathbb{P}[\hat{s} \geq (1 + 1)4\epsilon Mn] \leq e^{-\frac{4\epsilon Mn}{3}} = e^{-\frac{(1-\alpha)Mn}{6}} \leq e^{-\frac{1}{6}} \leq \frac{17}{20}.$$

So $\mathbb{P}[\hat{s} < (1 - \alpha)Mn] > \frac{3}{20}$.

Next, we consider the following sampling process with dataset $S = \{(x'_1, f(x'_1)), \dots, (x'_{M \times n}, f(x'_{M \times n}))\}$: choosing as many data $(x'_i, f(x'_i))$, $i \in [mn]$ such that $x'_i = x_1$ as possible to form the training set \bar{S} . Let $f_1, f_2 \in \mathcal{H}$ be two target functions whose existence is guaranteed by the non-trivial definition of \mathcal{H} and $f_1(x_1) = f_2(x_1)$, $f_1(x_2) = -f_2(x_2)$, and \mathcal{S} be the set of all datasets in $(\mathcal{X}, \mathcal{Y})^{(M \times n)}$ such that $\hat{s} < (1 - \alpha)Mn$.

Let $\mathcal{R}(h_s, f) = \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_1(x) \cap x \neq x_1]$, the following holds for these two target functions f_1 and f_2 :

$$\begin{aligned} \mathcal{R}(h_s, f_1) + \mathcal{R}(h_s, f_2) &= \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_1(x) \cap x \neq x_1] + \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_2(x) \cap x \neq x_1] \\ &= \mathbf{1}_{\mathcal{L}(\mathcal{F}(S))(x_1) \neq f_1(x_1)} \mathbb{P}(x_2) + \mathbf{1}_{\mathcal{L}(\mathcal{F}(S))(x_1) \neq f_2(x_2)} \mathbb{P}(x_1) \\ &= 4\epsilon. \end{aligned}$$

The above result hold in expectation since it holds for any $S \in \mathcal{S}$. Hence, there exists a target function $f \in \mathcal{H}$ such that $\mathbb{E}_{S \in \mathcal{S}} \mathcal{R}(h_s, f) \geq 2\epsilon$. Note $\mathcal{R}(h_s, f) \leq \mathbb{P}(x \neq x_1) = 4\epsilon$, then by decomposing the expectation into two parts we obtain:

$$\begin{aligned} 2\epsilon &\leq \mathbb{E}_{S \in \mathcal{S}} \mathcal{R}(h_s, f) = \sum_{S: \mathcal{R}(h_s, f) \geq \epsilon} \mathcal{R}(h_s, f) \mathbb{P}[\mathcal{R}(h_s, f)] + \sum_{S: \mathcal{R}(h_s, f) < \epsilon} \mathcal{R}(h_s, f) \mathbb{P}[\mathcal{R}(h_s, f)] \\ &\leq 4\epsilon \mathbb{P}_{S \in \mathcal{S}}[\mathcal{R}(h_s, f) \geq 4\epsilon] + \epsilon(1 - \mathbb{P}_{S \in \mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon]) \\ &= \epsilon + 3\epsilon \mathbb{P}_{S \in \mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon]. \end{aligned}$$

That is,

$$\mathbb{P}_{S \in \mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon] \geq \frac{1}{3}.$$

Note $\mathcal{R}(h_s, f) = \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_1(x) \cap x \neq x_1] \leq \mathcal{R}(\mathcal{L}(\mathcal{F}(S))) = \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_1(x)]$, then we have the final results:

$$\begin{aligned} \mathbb{P}_{S \sim P}[\mathcal{R}_P(\mathcal{L}(\mathcal{F}(S)), f) \geq \epsilon] &\geq \mathbb{P}_{S \sim P}[\mathcal{R}(h_s, f) \geq \epsilon] \\ &\geq \mathbb{P}_{S \in \mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon] \mathbb{P}[S \in \mathcal{S}] \\ &> \frac{1}{3} \frac{3}{20} = \frac{1}{20}. \end{aligned}$$

□

Theorem 2 (Generalization Error Bound for SA-FL). *For an SA-FL system with arbitrary system and data heterogeneity, if distributions P and Q satisfy Assumption 1 and are (α, β) -positively-related, then with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, it holds that*

$$\varepsilon_P(\hat{h}_Q^*) = \tilde{O} \left(\left(\frac{d_{\mathcal{H}}}{n_T + n_S} \right)^{\frac{1}{2-\beta Q}} + \left(\frac{d_{\mathcal{H}}}{n_T + n_S} \right)^{\frac{\beta}{2-\beta Q}} \right), \quad (1)$$

where $d_{\mathcal{H}}$ denotes the finite VC dimension for hypotheses class \mathcal{H} , and parameters $\{P, Q, n_T, n_S, \beta, \beta_Q\}$ are defined the same as before.

Proof.

$$\begin{aligned}\varepsilon_P(\hat{h}_Q^*) &= \mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) \\ &= [\mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) - (\mathcal{R}_Q(\hat{h}_Q^*) - \mathcal{R}_Q(h_Q^*))] + \mathcal{R}_Q(\hat{h}_Q^*) - \mathcal{R}_Q(h_Q^*) \\ &\leq |\varepsilon_P(\hat{h}_Q^*) - \varepsilon_Q(\hat{h}_Q^*)| + \varepsilon_Q(\hat{h}_Q^*) \\ &\leq \alpha \varepsilon_Q(\hat{h}_Q^*)^\beta + \varepsilon_Q(\hat{h}_Q^*).\end{aligned}$$

Combining with Lemma 1, the proof is complete.

Lemma 1 (Auxiliary Lemma (Massart & Nédélec, 2006; Koltchinskii, 2006; Hanneke & Kpotufe, 2019; 2020)). *For any $m \in \mathbb{N}$ and $\delta \in (0, 1)$, define $A(m, \delta) = \frac{d_{\mathcal{H}}}{m} \log(\frac{m}{d_{\mathcal{H}}} + \frac{1}{m} \log(\frac{1}{\delta}))$. With probability at least $1 - \delta$, $\forall h, \hat{h} \in \mathcal{H}$,*

$$\begin{aligned}\mathcal{R}(h) - \mathcal{R}(\hat{h}) &\leq \hat{\mathcal{R}}(h) - \hat{\mathcal{R}}(\hat{h}) + c\sqrt{\min\{\mathbb{P}_S(h \neq \hat{h}), \hat{\mathbb{P}}_S(h \neq \hat{h})\}A(m, \delta)} + cA(m, \delta), \\ \frac{1}{2}\mathbb{P}_S(h \neq \hat{h}) - cA(m, \delta) &\leq \hat{\mathbb{P}}_S(h \neq \hat{h}) \leq 2\mathbb{P}_S(h \neq \hat{h}) + cA(m, \delta), \\ \varepsilon_Q(\hat{h}_Q^*) &= [A(m, \delta)]^{\frac{1}{2-\beta_Q}},\end{aligned}$$

where $\mathbb{P}_S(\cdot) = \mathbb{E}[\hat{\mathbb{P}}_S(\cdot)]$, S is the i.i.d. dataset with size m drawn from distribution Q , $c \in (0, \infty)$ is a constant.

□

Theorem 3 (SA-FL Being No Worse Than Centralized Learning). *Consider an SA-FL system with arbitrary system and data heterogeneity. If Assumption 1 holds and additionally $\hat{\mathcal{R}}_P(\hat{h}_Q^*) \leq \hat{\mathcal{R}}_P(h_Q^*)$ and $\varepsilon_P(h_Q^*) = \mathcal{O}(\mathcal{A}(n_T, \delta))$, where $\mathcal{A}(n_T, \delta) = \frac{d_{\mathcal{H}}}{n_T} \log(\frac{n_T}{d_{\mathcal{H}}} + \frac{1}{n_T} \log(\frac{1}{\delta}))$, then with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, it holds that $\varepsilon_P(\hat{h}_Q^*) = \tilde{\mathcal{O}}\left((d_{\mathcal{H}}/n_T)^{\frac{1}{2-\beta_P}}\right)$. Other parameters are the same as defined in Theorem 2.*

Proof. Without loss of generality, we use c serve as a generic constant since we focus on the order in terms of the sample number and thus omit the constant factor.

$$\begin{aligned}\varepsilon_P(\hat{h}_Q^*) &= \mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) \\ &\leq \hat{\mathcal{R}}_P(\hat{h}_Q^*) - \hat{\mathcal{R}}_P(h_P^*) + c\sqrt{\min\{P(\hat{h}_Q^* \neq h_P^*), \hat{P}(\hat{h}_Q^* \neq h_P^*)\}A(n_T, \delta)} + cA(n_T, \delta) \\ &\leq c\sqrt{\varepsilon_P^{\beta_P}(\hat{h}_Q^*)A(n_T, \delta)} + cA(n_T, \delta).\end{aligned}$$

The first inequality is due to Lemma 1 and second inequality follows from Lemma 2 and Noise assumption 1. Then we have the following result, which completes the proof:

$$\varepsilon_P(\hat{h}_Q^*) \leq cA(n_T, \delta)^{\frac{1}{2-\beta_P}}.$$

□

Lemma 2. *If $\hat{\mathcal{R}}_P(\hat{h}_Q^*) \leq \hat{\mathcal{R}}_P(h_Q^*)$, with probability at least $1 - \delta$,*

$$\hat{\mathcal{R}}_P(\hat{h}_Q^*) - \hat{\mathcal{R}}_P(h_P^*) = \varepsilon_P(h_Q^*) + \mathcal{O}(A(n_T, \delta)).$$

Proof.

$$\begin{aligned}
\hat{\mathcal{R}}_P(\hat{h}_Q^*) - \hat{\mathcal{R}}_P(h_P^*) &\leq \hat{\mathcal{R}}_P(h_Q^*) - \hat{\mathcal{R}}_P(h_P^*) \\
&\leq \mathcal{R}_P(h_Q^*) - \mathcal{R}_P(h_P^*) + c\sqrt{\min\{P(h_Q^* \neq h_P^*), \hat{P}(h_Q^* \neq h_P^*)\}A(n_T, \delta)} + cA(n_T, \delta) \\
&= \varepsilon_P(h_Q^*) + \mathcal{O}(A(n_T, \delta)).
\end{aligned}$$

□

Theorem 4 (Convergence Rate for SAFARI). *Under Assumptions 2 and 3, let constant learning rate η satisfy $(\frac{1}{2} - 4LK\eta - 20K(L + 4KL^3\eta)\eta^2) > 0$. Then, the sequence $\{\mathbf{x}_r\}$ generated by the SAFARI algorithm satisfies:*

$$\begin{aligned}
\frac{1}{R} \sum_{t=0}^{R-1} \mathbb{E} \|\nabla F(\mathbf{x}_r)\|^2 &\leq \frac{1}{c} \left[\frac{F(\mathbf{x}_0) - F(\mathbf{x}^*)}{\eta KR} \right] + \frac{1}{c} [(5KL\eta^2 + 20K^2L^3\eta^3 + 2L\eta) \sigma^2] \\
&\quad + \frac{1}{c} \left[\left(\frac{1}{K^2} + \frac{L\eta}{K} \right) \frac{1}{R} \sum_{t=0}^{R-1} c_r^2 \right],
\end{aligned}$$

where c is a constant and \mathbf{x}^* denotes an optimal solution.

Proof. Let $\bar{\Delta}_r = \frac{1}{|S_r|} \sum_{i \in S_r} \hat{\Delta}_r^i$, $\bar{\mathbf{g}}_r = \Delta_r^0 + \frac{1}{|S_r|} \sum_{i \in S_r} \hat{\Delta}_r^i = \Delta_r^0 + \bar{\Delta}_r$.

Let $\mathbb{E}_r[\cdot]$ denote the conditional expectation conditioned on \mathbf{x}_r , which is averaged over all realizations of the random dataset T , we have

$$\begin{aligned}
\mathbb{E}_r[F(\mathbf{x}_{r+1})] &\leq F(\mathbf{x}_r) + \langle \nabla F(\mathbf{x}_r), \mathbb{E}_r[\mathbf{x}_{r+1} - \mathbf{x}_r] \rangle + \frac{L}{2} \mathbb{E}_r[\|\mathbf{x}_{r+1} - \mathbf{x}_r\|^2] \\
&= F(\mathbf{x}_r) + \langle \nabla F(\mathbf{x}_r), \eta \mathbb{E}_r \bar{\mathbf{g}}_r \rangle + \frac{L}{2} \eta^2 \mathbb{E}_r[\|\bar{\mathbf{g}}_r\|^2] \\
&= F(\mathbf{x}_r) - \eta K \|\nabla F(\mathbf{x}_r)\|^2 + \underbrace{\langle \nabla F(\mathbf{x}_r), \eta K \nabla F(\mathbf{x}_r) + \eta \mathbb{E}_r [\Delta_r^0 + \bar{\Delta}_r] \rangle}_{A_1} + \underbrace{\frac{L}{2} \eta^2 \mathbb{E}_r [\|\Delta_r^0 + \bar{\Delta}_r\|^2]}_{A_2}.
\end{aligned}$$

The we can bound A_1 and A_2 separately as follows.

$$\begin{aligned}
A_1 &= \langle \nabla F(\mathbf{x}_r), \eta K \nabla F(\mathbf{x}_r) + \eta \mathbb{E}_r [\Delta_r^0 + \bar{\Delta}_r] \rangle = \eta K \langle \nabla F(\mathbf{x}_r), \nabla F(\mathbf{x}_r) + \frac{1}{K} \mathbb{E}_r [\Delta_r^0 + \bar{\Delta}_r] \rangle \\
&\leq \frac{1}{2} \eta K \|\nabla F(\mathbf{x}_r)\|^2 + \frac{1}{2} \eta K \left\| \nabla F(\mathbf{x}_r) + \frac{1}{K} \mathbb{E}_r [\Delta_r^0 + \bar{\Delta}_r] \right\|^2
\end{aligned}$$

Note that $\Delta_r^0 = -\sum_{k=0}^{K-1} \nabla F(\mathbf{x}_{r,k}^0, \xi_{r,k}^0)$. We have

$$\begin{aligned}
\frac{1}{2} \eta K \left\| \nabla F(\mathbf{x}_r) + \frac{1}{K} \mathbb{E}_r [\Delta_r^0 + \bar{\Delta}_r] \right\|^2 &\leq \eta K \left\| \nabla F(\mathbf{x}_r) + \frac{1}{K} \mathbb{E}_r [\Delta_r^0] \right\|^2 + \eta K \left\| \frac{1}{K} \mathbb{E}_r [\bar{\Delta}_r] \right\|^2 \\
&\leq \eta K \left\| \nabla F(\mathbf{x}_r) - \mathbb{E}_r \left[\frac{1}{K} \sum_{k=0}^{K-1} \nabla F(\mathbf{x}_{r,k}^0, \xi_{r,k}^0) \right] \right\|^2 + \frac{\eta}{K} \mathbb{E}_r \|\bar{\Delta}_r\|^2 \\
&= \eta K \left\| \nabla F(\mathbf{x}_r) - \frac{1}{K} \sum_{k=0}^{K-1} \nabla F(\mathbf{x}_{r,k}^0) \right\|^2 + \frac{\eta}{K} \mathbb{E}_r \left\| \frac{1}{|S_r|} \sum_{i \in S_r} \hat{\Delta}_r^i \right\|^2 \\
&\leq \eta \sum_{k=0}^{K-1} \|\nabla F(\mathbf{x}_r) - \nabla F(\mathbf{x}_{r,k}^0)\|^2 + \eta \frac{1}{|S_r|} \mathbb{E}_r \sum_{i \in S_r} \|\hat{\Delta}_r^i\|^2 \\
&\leq \eta L \sum_{k=0}^{K-1} \|\mathbf{x}_r - \mathbf{x}_{r,k}^0\|^2 + \frac{\eta}{K} c_r^2.
\end{aligned}$$

Note the first equality follows from the fact that $\mathbb{E}_r[\cdot]$ is over all realizations of the random dataset T uniformly and independently sampled from P , rather than a single realization (i.e., a fixed dataset). This implies that $\xi_{r,k}^0$ is an i.i.d. data sample from a random dataset T which is uniformly sampled from P .

So, we can bound A_1 as following:

$$A_1 \leq \frac{1}{2}\eta K \|\nabla F(\mathbf{x}_r)\|^2 + \eta L \sum_{k=0}^{K-1} \|\mathbf{x}_r - \mathbf{x}_{r,k}^0\|^2 + \frac{\eta}{K} c_r^2.$$

$$A_2 = \frac{L}{2} \eta^2 \mathbb{E} [\|\Delta_r^0 + \bar{\Delta}_r\|^2] \leq L \eta^2 \mathbb{E}_r [\|\Delta_r^0\|^2] + L \eta^2 \mathbb{E}_r [\|\bar{\Delta}_r\|^2].$$

$$\begin{aligned} \mathbb{E}_r [\|\Delta_r^0\|^2] &= \mathbb{E}_r \left\| \sum_{k=0}^{K-1} \nabla F(\mathbf{x}_{r,k}^0, \xi_{r,k}^0) \right\|^2 \\ &\leq 2 \left\| \sum_{k=0}^{K-1} \nabla F(\mathbf{x}_{r,k}^0) \right\|^2 + 2K\sigma^2 \\ &\leq 2 \left\| \sum_{k=0}^{K-1} [\nabla F(\mathbf{x}_{r,k}^0) - \nabla F(\mathbf{x}_r) + \nabla F(\mathbf{x}_r)] \right\|^2 + 2K\sigma^2 \\ &\leq 4K \sum_{k=0}^{K-1} [\|\nabla F(\mathbf{x}_{r,k}^0) - \nabla F(\mathbf{x}_r)\|^2 + \|\nabla F(\mathbf{x}_r)\|^2] + 2K\sigma^2 \\ &\leq 2K\sigma^2 + 4KL^2 \sum_{k=0}^{K-1} \|\mathbf{x}_{r,k}^0 - \mathbf{x}_r\|^2 + 4K^2 \|\nabla F(\mathbf{x}_r)\|^2, \end{aligned}$$

where the first inequality is due to assumption 1 and $\{\nabla F(\mathbf{x}_{r,k}^0, \xi_{r,k}^0) - \nabla F(\mathbf{x}_{r,k}^0)\}$ form a martingale difference sequence (see Lemma 4 in (Karimireddy et al., 2020)).

Hence, we can bound A_2 as following:

$$A_2 \leq 2KL\eta^2\sigma^2 + 4KL^3\eta^2 \sum_{k=0}^{K-1} \|\mathbf{x}_{r,k}^0 - \mathbf{x}_r\|^2 + 4LK^2\eta^2 \|\nabla F(\mathbf{x}_r)\|^2 + L\eta^2 c_r^2.$$

By plugging the bound of A_1 and A_2 into the smoothness inequality and taking full expectation, we have:

$$\begin{aligned} \mathbb{E}[F(\mathbf{x}_{r+1}) - F(\mathbf{x}_r)] &\leq \mathbb{E} [-\eta K \|\nabla F(\mathbf{x}_r)\|^2 + A_1 + A_2] \\ &\leq \mathbb{E} \left[-\eta K \left(\frac{1}{2} - 4LK\eta \right) \|\nabla F(\mathbf{x}_r)\|^2 + (\eta L + 4KL^3\eta^2) \sum_{k=0}^{K-1} \mathbb{E}_r [\|\mathbf{x}_{r,k}^0 - \mathbf{x}_r\|^2] + \left(\frac{\eta}{K} + L\eta^2 \right) c_r^2 + 2KL\eta^2\sigma^2 \right]. \end{aligned}$$

For the server, we have the following results for the norm of parameter changes for one local computation:

$$\begin{aligned} \mathbb{E}_r [\|\mathbf{x}_{r,k}^0 - \mathbf{x}_r\|^2] &= \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r - \eta g_{t,k-1}^0\|^2] \\ &= \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r - \eta \nabla F(\mathbf{x}_{t,k-1}^0)\|^2] + \mathbb{E}_r [\|\eta (g_{t,k-1}^0 - \nabla F(\mathbf{x}_{t,k-1}^0))\|^2] \\ &= (1 + \frac{1}{2K-1}) \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r\|^2] + \mathbb{E}_r [\|\eta (g_{t,k-1}^0 - \nabla F(\mathbf{x}_{t,k-1}^0))\|^2] \\ &\quad + 2K \mathbb{E}_r [\|\eta \nabla F(\mathbf{x}_{t,k-1}^0) - \eta \nabla F(\mathbf{x}_r) + \eta \nabla F(\mathbf{x}_r)\|^2] \\ &= (1 + \frac{1}{2K-1}) \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r\|^2] + \mathbb{E}_r [\|\eta (g_{t,k-1}^0 - \nabla F(\mathbf{x}_{t,k-1}^0))\|^2] \end{aligned}$$

$$\begin{aligned}
& + 4K\eta^2 \|\nabla F(\mathbf{x}_{t,k-1}^0) - \nabla F(\mathbf{x}_r)\|^2 + 4K\eta^2 \|\nabla F(\mathbf{x}_r)\|^2 \\
& \leq (1 + \frac{1}{2K-1} + 4KL^2\eta^2) \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r\|^2] + \eta^2\sigma^2 + 4K\eta^2 \|\nabla F(\mathbf{x}_r)\|^2 \\
& \leq (1 + \frac{1}{K-1}) \mathbb{E}_r [\|\mathbf{x}_{t,k-1}^0 - \mathbf{x}_r\|^2] + \eta^2\sigma^2 + 4K\eta^2 \|\nabla F(\mathbf{x}_r)\|^2.
\end{aligned}$$

Unrolling the recursion, we obtain the following:

$$\begin{aligned}
\mathbb{E}_r [\|\mathbf{x}_{r,k}^0 - \mathbf{x}_r\|^2] &= \sum_{p=0}^{k-1} (1 + \frac{1}{K-1})^p (\eta^2\sigma^2 + 4K\eta^2 \|\nabla F(\mathbf{x}_r)\|^2) \\
&\leq (K-1) \left[\left(1 + \frac{1}{K-1}\right)^K - 1 \right] (\eta^2\sigma^2 + 4K\eta^2 \|\nabla F(\mathbf{x}_r)\|^2) \\
&\leq 5K\eta^2\sigma^2 + 20K^2\eta^2 \|\nabla F(\mathbf{x}_r)\|^2.
\end{aligned}$$

Putting the pieces together, we obtain

$$\begin{aligned}
& \mathbb{E} [F(\mathbf{x}_{r+1}) - F(\mathbf{x}_r)] \\
& \leq \mathbb{E} \left[-\eta K \left(\frac{1}{2} - 4LK\eta - 20K(L + 4KL^3\eta)\eta^2 \right) \|\nabla F(\mathbf{x}_r)\|^2 \right] \\
& \quad + \mathbb{E} \left[\eta K (5KL\eta^2 + 20K^2L^3\eta^3 + 2L\eta) \sigma^2 + \left(\frac{\eta}{K} + L\eta^2 \right) c_r^2 \right] \\
& \leq \mathbb{E} \left[-c\eta K \|\nabla F(\mathbf{x}_r)\|^2 + \eta K (5KL\eta^2 + 20K^2L^3\eta^3 + 2L\eta) \sigma^2 + \left(\frac{\eta}{K} + L\eta^2 \right) c_r^2 \right].
\end{aligned}$$

The last inequality follows from that there exist such constant c if $(\frac{1}{2} - 4LK\eta - 20K(L + 4KL^3\eta)\eta^2) > 0$.

Summing over $r = 0$ to $R-1$, we have

$$\begin{aligned}
\frac{1}{R} \sum_{t=0}^{R-1} \mathbb{E} \|\nabla F(\mathbf{x}_r)\|^2 &\leq \frac{1}{c} \left[\frac{F(\mathbf{x}_0) - F(\mathbf{x}^*)}{\eta KR} + (5KL\eta^2 + 20K^2L^3\eta^3 + 2L\eta) \sigma^2 \right. \\
&\quad \left. + \left(\frac{1}{K^2} + \frac{L\eta}{K} \right) \frac{1}{R} \sum_{t=0}^{R-1} c_r^2 \right]
\end{aligned}$$

□

B EXPERIMENTS

In this section, we provide the details of the numerical experiments and some additional experimental results.

B.1 MODELS AND DATASETS

We test the SAFARI algorithm by running two models on two different types of datasets, including 1) multinomial logistic regression (LR) on MNIST, and 2) convolutional neural network (CNN) on CIFAR-10. Both datasets are chose from a previous FL paper (McMahan et al., 2017), and they are now widely used as benchmarks for FL research (Yang et al., 2021b; Li et al., 2020b).

MNIST and CIFAR-10 have ten classes of images separately. In order to impose the heterogeneity of the data, we partition the dataset according to the number of classes (p) that each client contains. We distribute these data to $M = 10$ clients, and each client only has a certain number of classes. Specifically, each client randomly selects p classes of images and then evenly samples training and test data-points within these p classes of images without replacement. For example, if $p = 2$, each client only samples training and test data-points within two classes of images, which causes the

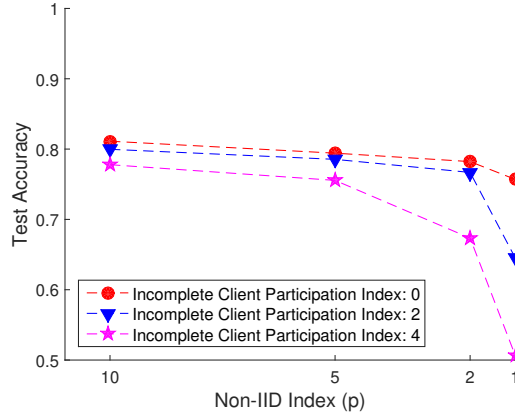


Figure 3: Test Accuracy of FedAvg on CIFAR-10 with incomplete client participation. Larger incomplete client participation index means less clients participate in the training, and smaller non-i.i.d. index means the data across clients is more heterogeneous.

heterogeneity among different clients. If $p = 10$, each client contains training and test samples that selects from ten classes. This situation is almost the same as i.i.d. case. Hence, the number of classes (p) in each client’s local dataset can be used to represent the level of non-i.i.d. qualitatively. In addition, to mimic incomplete client participation, we enforce s clients to be exempt from participation, where the index s can be used to represent the degree of incomplete client participation. Specifically, we assume there are $M = 10$ clients in total, and $m = 5$ clients participate in each communication round. These clients are uniformly sampled from $M - s$ clients. Larger incomplete client participation index s means less clients participate in the training.

For both MNIST and CIFAR-10, the learning rate is 0.1, and the local epoch is 1. For MNIST, the batch size is 64, and the total communication round is 150. For CIFAR-10, the batch size is 500, and the total communication round is 4000. To simulate the data heterogeneity, we use $p = [10, 5, 2, 1]$ as a proxy to represent the degree of non-i.i.d. on MNIST and CIFAR-10 datasets. To emulate the effect of incomplete client participation, we set $s = [0, 2, 4]$ to represent the degree of incomplete client participation for the SAFARI algorithm, the FedAvg algorithm, and the SGD algorithm. Last two algorithms are employed as the baselines to compare with our algorithm. The hyper-parameter c_t in the SAFARI algorithm is set to 0.1 both on MNIST and CIFAR-10. To compare the effect of the collaboration from server, we add $[50, 100, 500, 1000]$ data to the server’s side for MNIST and $[500, 1000, 5000]$ for CIFAR-10.

B.2 ADDITIONAL EXPERIMENTAL RESULTS

In Figure B.2, we show the test accuracy of FedAvg algorithm on CIFAR-10 for different Non-IID index p and incomplete client participation index s . In the case of $p = 10$, the test accuracy of $s = 4$ and $s = 0$ is not much different whereas the test accuracy of $s = 4$ is 25% lower than that of $s = 1$ in the case of $p = 1$. This finding on CIFAR-10 further support our first observation in Section 5. Incomplete client participation has no impact on the performance for nearly homogeneous data, but it causes catastrophic performance degradation for highly Non-IID data.

In Figure 4, we show the test accuracy of the SAFARI algorithm, the FedAvg algorithm, and the SGD algorithm on MNIST for incomplete client participation $s = 4$ and different Non-IID index p . The evidences of the observations in Section 5 are provided visually as follows:

- Compared to FedAvg in the case of $p = 1$ (see Figure 4(d)), with only 50 data at server’s side (0.1% of the total training data), there is a non-negligible increase of test accuracy for our SAFARI algorithm. This increase increases as more data is added to the server’s side.
- In nearly homogeneous case when $p = 5$ or $p = 10$ (see Figure 4(a) and 4(b)), there is actually no improvement of the test accuracy with these auxiliary data added to the server’s side, comparing SAFARI with FedAvg.

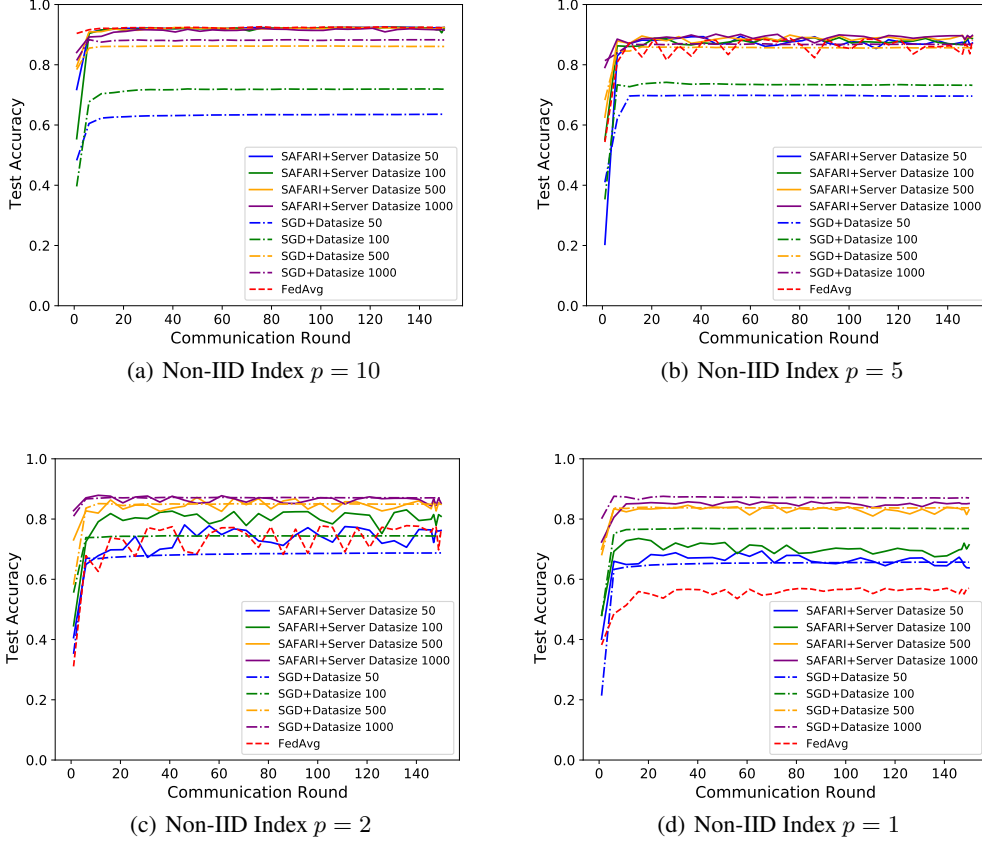


Figure 4: Test accuracy of SAFARI , FedAvg, and SGD algorithm on MNIST with incomplete client participation $s = 4$ and different Non-IID index p . Smaller p means the data across clients is more heterogeneous.

Table 3: Test accuracy improvement (%) for SAFARI compared with FedAvg on CIFAR-10 with incomplete client participation $s = 4$. ‘-’ means no statistical difference within 2% error bar.

SERVER DATASIZE	NON-IID INDEX (p)			
	10	5	2	1
500	-	-	-	-
1000	-	-	-	3.55
5000	-	-	5.45	16.08

- Compared SAFARI with SGD (for centralized learning solely on server’s data) in nearly homogeneous case when $p = 5$ or $p = 10$ (see Figure 4(a) and 4(b)), the collaborations from clients significantly improves the performance, especially with less data on the server’s side.
- In highly heterogeneous case when $p = 2$ or $p = 1$ (see Figure 4(c) and 4(d)), it shows no obvious improvement from the collaboration of clients comparing SAFARI to SGD.

In Table 3, we show the comparison between our SAFARI algorithm and FedAvg algorithm on CIFAR-10 for incomplete client participation $s = 4$. The observations in Section 5 are further illustrated: 1) There is non-negligible increase of the test accuracy for SAFARI algorithm with small amount of auxiliary data at server’s side. With 5000 data at server’s side, the test accuracy increases by 16.08%. 2) There is actually no improvement with these auxiliary data for nearly homogeneous case (e.g., $p = 10$ or $p = 5$), which is denoted by ‘-’ in the table.

Table 4: Test accuracy improvement (%) of SAFARI under incomplete client participation $s = 4$ compared with SGD in centralized learning on CIFAR-10. Smaller Non-IID index means the data across clients is more heterogeneous.

SERVER DATASIZE	NON-IID INDEX (p)			
	10	5	2	1
500	35.67	33.48	27.60	10.77
1000	31.23	28.46	22.36	7.62
5000	13.99	11.11	7.88	3.40

In Table 4, we show the difference between our SAFARI algorithm and SGD, which is for centralized learning solely on server’s data, for incomplete client participation $s = 4$ on CIFAR-10. When the size of data on server’s side is small, the collaborations from clients significantly improve the performance of the SAFARI algorithm. Even in the highly heterogeneous case when $p = 1$, the test accuracy can be improved by 10% for only 500 data on the server’s side (0.8% of the total training data). This observation further validates our theoretical analysis in Theorem 2.