Understanding Server-Assisted Federated Learning in the Presence of Incomplete Client Participation

Haibo Yang [1] **Peiwen Qiu** [2] **Prashant Khanduri** [3] **Minghong Fang** [4] **Jia Liu** [2]

Abstract

Existing works in federated learning (FL) often assume an ideal system with either full client or uniformly distributed client participation. However, in practice, it has been observed that some clients may never participate in FL training (aka incomplete client participation) due to a myriad of system heterogeneity factors. A popular approach to mitigate impacts of incomplete client participation is the server-assisted federated learning (SA-FL) framework, where the server is equipped with an auxiliary dataset. However, despite SA-FL has been empirically shown to be effective in addressing the incomplete client participation problem, there remains a lack of theoretical understanding for SA-FL. Meanwhile, the ramifications of incomplete client participation in conventional FL are also poorly understood. These theoretical gaps motivate us to rigorously investigate SA-FL. Toward this end, we first show that conventional FL is *not* PAC-learnable under incomplete client participation in the worst case. Then, we show that the PAC-learnability of FL with incomplete client participation can indeed be revived by SA-FL, which theoretically justifies the use of SA-FL for the first time. Lastly, to provide practical guidance for SA-FL training under *incomplete client participation*, we propose the SAFARI (server-assisted federated averaging) algorithm that enjoys the same linear convergence speedup guarantees as classic FL with ideal client participation assumptions, offering the first SA-FL algorithm with convergence guarantee. Extensive experiments on different datasets show SAFARI significantly improves the performance under incomplete client participation.

1. Introduction

Since the seminal work by McMahan et al. (2017), federated learning (FL) has emerged as a powerful distributed learning paradigm that enables a large number of clients (e.g., edge devices) to collaboratively train a model under a central server's coordination. However, as FL gaining popularity, it has also become apparent that FL faces a key challenge unseen in traditional distributed learning in data-center settings – system heterogeneity. Generally speaking, system heterogeneity in FL is caused by the massively different computation and communication capabilities at each client (computational power, communication capacity, drop-out rate, etc.). Studies have shown that system heterogeneity can significantly impact client participation in a highly non-trivial fashion and render *incomplete client participation*, which severely degrades the learning performance (Bonawitz et al., 2019; Yang et al., 2021a). For example, it is shown in (Yang et al., 2021a) that more than 30% clients never participate in FL, while only 30% of the clients contribute to 81% of the total computation even if the server uniformly samples the clients. Exacerbating the problem is the fact that clients' status could be unstable and time-varying due to the aforementioned computation/communication constraints. This situation sharply contrasts with existing works on FL with partial client participation, which often assume that clients engage based on a known random process (Karimireddy et al., 2020; Malinovsky et al., 2023; Cho et al., 2023).

To mitigate the impact of incomplete client participation, one approach called *server-assisted federated learning* (SA-FL) has been widely adopted in real-world FL systems in recent years (see, e.g., (Zhao et al., 2018; Wang et al., 2021b)). The basic idea of SA-FL is to equip the server with a small auxiliary dataset sampled from population distribution, so that the distribution deviation induced by incomplete client participation can be corrected. Nonetheless, while SA-FL has empirically demonstrated its considerable efficacy in addressing incomplete client participation problem in practice,

[1]Department of Computing and Information Sciences Ph.D., Rochester Institute of Technology, Rochester, NY, USA [2]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA [3]Department of Computer Science, Wayne State University, Detroit, MI, USA [4]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. Correspondence to: Haibo Yang <hbycis@rit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

there remains *a lack of theoretical understanding* for SA-FL. This motivates us to rigorously investigate the efficacy of SA-FL in the presence of incomplete client participation.

Somewhat counterintuitively, to understand SA-FL, one must first fully understand the impact of incomplete client participation on conventional FL. In other words, we need to first answer the following fundamental question:

(Q1): What are the impacts of incomplete client participation on conventional FL learning performance?

Upon answering this question, the next important follow-up question regarding SA-FL is:

(Q2): What benefits could SA-FL bring and how could we theoretically characterize them?

Also, just knowing the benefits of SA-FL is not sufficient to provide guidelines on how to use server-side data in designing training algorithms with convergence guarantees. Therefore, our third fundamental question for SA-FL is:

(Q3): Is it possible to develop SA-FL training algorithms with provable convergence rates that can match the state-of-the-art rates in conventional FL?

Answering these three questions constitutes the rest of this paper, where we address Q1 and Q2 through the lens of PAC (probably approximately correct) learning, while resolving Q3 by proposing a provably convergent SA-FL algorithm. Our major contributions are summarized as follows:

- By establishing a *worst-case* generalization error lower bound, we rigorously show that classic FL is *not* PAC-learnable under incomplete client participation. In other words, no learning algorithm can approach zero generalization error with incomplete client participation for classic FL even in the limit of infinitely many data samples. This insight, though being negative, warrants the necessity of developing new algorithmic techniques and system architectures (e.g., SA-FL) to modify the classic FL framework to mitigate incomplete client participation.
- We prove a new generalization error bound to show that SA-FL can indeed *revive the PAC learnability of FL* with incomplete client participation. We note that this bound could reach zero asymptotically as the number data samples increases. This is much stronger than previous results in domain adaptation with non-vanishing small error (see Section 2 for details).
- To ensure that SA-FL is provably convergent in training, we propose a new training algorithm for SA-FL called SAFARI (server-assisted federated averaging). By carefully designing the server-client update coordination, we show that SAFARI achieves an $\mathcal{O}(1/\sqrt{mkR})$ convergence rate in non-convex functions and $\tilde{\mathcal{O}}(\frac{1}{R})$ in strongly-convex functions, matching the convergence rates of state-of-the-art classic FL algorithms (Li et al., 2020b; Yang et al., 2021b). We also conduct extensive experiments to demonstrate the effectiveness of our SAFARI algorithm.

2. Related Work

In this section, we provide an overview on two lines of closely related research, namely (i) FL with partial client participation and (ii) domain adaptation.

1) Partial Client Participation in Federated Learning: Since The seminal FedAvg algorithm (McMahan et al., 2017), there have been many follow-ups (e.g., (Li et al., 2020a; Wang et al., 2020; Zhang et al., 2020; Acar et al., 2021; Karimireddy et al., 2020; Luo et al., 2021; Mitra et al., 2021; Karimireddy et al., 2021; Khanduri et al., 2021; Murata & Suzuki, 2021; Avdiukhin & Kasiviswanathan, 2021; Yang et al., 2021b; Grudzień et al., 2023; Condat et al., 2023; Mishchenko et al., 2022) and so on) on addressing the data heterogeneity challenge in FL. However, most of these works are based on the full or uniform (i.e., sampling clients uniformly at random) client participation assumption.

A related line of works in FL different from full/uniform client participation focuses on *proactively creating* flexible client participation (see, e.g., (Xie et al., 2019; Ruan et al., 2021; Gu et al., 2021; Avdiukhin & Kasiviswanathan, 2021; Yang et al., 2022; Wang & Ji, 2022; Koloskova et al., 2022)). The main idea here is to allow asynchronous communication or fixed participation pattern (e.g., given probability) for clients to flexibly participate in training. Existing works in this area often require extra assumptions, such as bounded delay (Ruan et al., 2021; Gu et al., 2021; Yang et al., 2022; Koloskova et al., 2022) and identical computation rate (Avdiukhin & Kasiviswanathan, 2021). Moreover, several papers explore unique scenarios of client participation. For instance, (Chen et al., 2020) selects the optimal client subset to minimize gradient estimation errors. The studies by (Malinovsky et al., 2023) and (Cho et al., 2023) investigated cyclic client participation. Additionally, in (Wang & Ji, 2023), optimal weights for each client are determined based on the estimated probabilities of their participation. In contrast, this paper addresses a more practical worst-case scenario in FL – "incomplete client participation." This phenomenon may arise from various heterogeneous factors, as discussed in Section 1.

2) Domain Adaptation: Since incomplete client participation induces a gap between the dataset distribution used for FL training and the true data population distribution across all clients, our work is also related to the field of domain adaptation. Domain adaptation focuses on the learnability of a model trained in one source domain but applied to a different and related target domain. The basic

approach is to quantify the error in terms of the source domain plus the distance between source and target domains. Specifically, let P and Q be the target and source distributions, respectively. Then, the generalization error is expressed as $\mathcal{O}(\mathcal{A}(n_Q)) + dist(P, Q)$, where $\mathcal{A}(n_Q)$ is an upper bound of the error dependent on the total number of samples in Q. Widely-used distance measures include $d_{\mathcal{A}}$-divergence (Ben-David et al., 2010; David et al., 2010) and $\mathcal{Y}$-discrepancy (Mansour et al., 2009; Mohri & Medina, 2012). We note, however, that results in domain adaptation is not directly applicable in FL with incomplete client participation, since doing so yields an overly pessimistic bound. Specifically, the error based on domain adaptation remains non-zero for asymptotically small distance $dist(P, Q)$ between P and Q even with infinite many samples in n_Q (i.e., $\mathcal{A}(n_Q) \to 0$). In this paper, rather than directly using results from domain adaptation, we establish a much *sharper* upper bound (see Section 3). A closely related work is (Hanneke & Kpotufe, 2019), which proposed a new notion of discrepancy between source and target distributions. However, this work considers *non-overlapping* support between P and Q, while we focus on *overlapping* support naturally implied by FL (see Fig. 1 in Section 3.2).

3. PAC-Learnability of Federated Learning with Incomplete Client Participation

In this section, we first focus on understanding the impacts of incomplete client participation on conventional FL in Section 3.1. This will also pave the way for studying SA-FL later in Section 3.2. In what follows, we start with FL formulation and some definitions in statistical learning that are necessary to formulate and prove our main results.

The goal of an M-client FL system is to minimize the following loss function $F(\mathbf{x}) = \mathbb{E}_{i\sim\mathcal{P}}[F_i(\mathbf{x})]$, where $F_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi\sim P_i}[f_i(\mathbf{x}, \xi)]$. Here, $\mathcal{P}$ represents the distribution of the entire client population, $\mathbf{x} \in \mathbb{R}^d$ is the model parameter, $F_i(\mathbf{x})$ represents the local loss function at client i, and P_i is the underlying distribution of the local dataset at client i. In general, due to data heterogeneity, $P_i \neq P_j$ if $i \neq j$. However, the loss function $F(\mathbf{x})$ or full gradient $\nabla F(\mathbf{x})$ can not be directly computed since the exact distribution of data is unknown in general. Instead, one often considers the following empirical risk minimization (ERM) problem in the finite-sum form based on empirical risk $\hat{F}(\mathbf{x})$:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \hat{F}(\mathbf{x}) \triangleq \sum_{i\in[M]} \alpha_i \hat{F}_i(\mathbf{x}), \hat{F}_i(\mathbf{x}) \triangleq \sum_{\xi\in S_i} f_i(\mathbf{x}, \xi),$$

where S_i is a local dataset at client i with cardinality $|S_i|$, whose samples are independently and identically sampled from distribution P_i, and $\alpha_i = |S_i|/(\sum_{j\in[M]} |S_j|)$ (hence $\sum_{i\in[M]} \alpha_i = 1$). For simplicity, we consider the balanced dataset case: $\alpha_i = 1/M, \forall i \in [M]$. Next, we state several definitions from statistical learning (Mohri et al., 2018).

Definition 1 (Generalization and Empirical Errors)**.** *Given a hypothesis $h \in \mathcal{H}$, a target concept f, an underlying distribution $\mathcal{D}$ and a dataset S i.i.d. sampled from $\mathcal{D}$ ($S \sim \mathcal{D}$), the generalization error and empirical error of h are defined as follows: $\mathcal{R}_{\mathcal{D}}(h, f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} l(h(x), f(x))$ and $\hat{\mathcal{R}}_D(h, f) = (1/|S|) \sum_{i\in S} l(h(x_i), f(x_i))$, where $l(\cdot)$ is a valid loss function.*

For simplicity, we will use $\mathcal{R}_{\mathcal{D}}(h)$ and $\hat{\mathcal{R}}_D(h)$ for generalization and empirical errors and omit target concept f.

Definition 2 (Optimal Hypothesis)**.** *We define $h^*_{\mathcal{D}} = \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{R}_{\mathcal{D}}(h)$ and $\hat{h}^*_{\mathcal{D}} = \operatorname{argmin}_{h\in\mathcal{H}} \hat{\mathcal{R}}_{\mathcal{D}}(h)$.*

Definition 3 (Excess Error)**.** *The excess error and excess empirical error are defined as $\varepsilon_{\mathcal{D}}(h) = \mathcal{R}_{\mathcal{D}}(h) - \mathcal{R}_{\mathcal{D}}(h^*_{\mathcal{D}})$, and $\hat{\varepsilon}_{\mathcal{D}}(h) = \hat{\mathcal{R}}_{\mathcal{D}}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(\hat{h}^*_{\mathcal{D}})$, respectively.*

3.1. Conventional Federated Learning with Incomplete Client Participation

With the above notations, we now study conventional FL with incomplete client participation (Q1). Consider an FL system with M clients in total. We let P denote the underlying joint distribution of the entire system, which can be decomposed into the summation of the local distributions at each client, i.e., $P = \sum_{i\in[M]} \lambda_i P_i$, where $\lambda_i > 0$ and $\sum_{i\in[M]} \lambda_i = 1$. We assume that each client i has n training samples i.i.d. drawn from P_i, i.e., $|S_i| = n, \forall i \in [M]$. Then, $S = \{(x_i, y_i), i \in [M \times n]\}$ can be viewed as the dataset i.i.d. sampled from the joint distribution P. We consider an incomplete client participation setting, where $m \in [0, M)$ clients participate in the FL training as a result of some client sampling/participation process $\mathcal{F}$. We let $\mathcal{F}(S)$ represent the data ensemble actually used in training and $\mathcal{D}$ denote the underlying distribution corresponding to $\mathcal{F}(S)$. For convenience, we define the notion $\omega = \frac{m}{M}$ as the *FL system capacity* (i.e., only m clients participate in the training). For FL with incomplete client participation, we establish the following fundamental performance limit of any FL learner in general. For simplicity, we use binary classification with zero-one loss here, but it is already sufficient to establish the PAC learnability lower limit.

Definition 4. *A concept class $\mathcal{C}$ is said to be PAC learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot, \cdot, \cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, and for any distribution $\mathcal{D}$ over the instance space $\mathcal{X}$, the algorithm, with probability at least $1 - \delta$, outputs a hypothesis h such that: $\mathbb{P}_{S\sim D}(R(h_S) \leq \epsilon) \geq 1 - \delta$, where $R(h)$ denotes the generalization error of hypothesis h returned by the algorithm. When such an algorithm $\mathcal{A}$ exists, it is called a PAC-learning algorithm for $\mathcal{C}$.*

In plain language, being PAC-learnable requires the hypothesis (or the model) returned by the algorithm after observing

enough number of data samples is approximately correct (error at most ϵ) with high probability (at least $1-\delta$).

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

Proof Sketch. The proof is based on the method of induced distributions in (Bshouty et al., 2002; Mohri et al., 2018; Konstantinov et al., 2020). We first show that the learnability of an FL system is equivalent to that of a system that arbitrarily selects mn out of Mn samples in the centralized learning. Then, for any learning algorithm, there exists a distribution P such that dataset $\mathcal{F}(S)$ resulting from incomplete participation and seen by the algorithm is always distributed identically for any target functions. Due to space limitation, we relegate the full proof to appendix. □

Given the system capacity $\omega\in(0,1)$, the above theorem characterizes the worst-case scenario for FL with incomplete client participation. It says that for any learner (i.e., algorithm) $\mathcal{L}$, there exist a bad client participation process $\mathcal{F}$ and distributions $P_i, i\in[M]$ over target function f, for which the error of the hypotheses returned by $\mathcal{L}$ is constant with non-zero probability. In other words, FL with incomplete client participation is *not PAC-learnable*. One interesting observation here is that the lower bound is *independent* of the number of samples per client n. This indicates that even if each client has *infinitely many* samples ($n\to\infty$), it is impossible to have a zero-generation-error learner under the incomplete client participation (i.e., $\omega\in(0,1)$). Note that this fundamental result relies on two conditions: *heterogeneous* dataset and *arbitrary* client participation. Under these two conditions, there exists a worst-case scenario where the underlying distribution $\mathcal{D}$ of the participating data $S_{\mathcal{D}}=\mathcal{F}(S)$ deviates from the ground truth P, thus yielding a non-vanishing error.

This result sheds light on system and algorithm design for FL. That is, how to motivate client participation in FL effectively and efficiently: the participating client's data should be comprehensive enough to model the complexity of the joint distribution P to close the gap between $\mathcal{D}$ and P. Note that this result is not contradictory to previous works where the convergence of FedAvg is guaranteed, since this theorem is not applicable for homogeneous (i.i.d.) datasets or uniformly random client participation. As mentioned earlier, most of the existing works rely on at least one of these two assumptions. However, none of these two assumptions hold for conventional FL with incomplete client participation in practice. In addition to system heterogeneity, other factors such as Byzantine attackers could also render incomplete client participation. For example, even for full client participation in FL, if part of the clients are Byzantine attackers, the impossibility theorem also applies. Thus, our impossibility theorem also justifies the empirical use of server-assisted federated learning (i.e., FL with server-side auxiliary data) to build trust (Cao et al., 2021).

3.2. The PAC-Learnability of Server-Assisted Federated Learning (SA-FL)

The intuition of SA-FL is to utilize a dataset T i.i.d. sampled from distribution P with cardinality $|T|=n_T$ as a vehicle to correct potential distribution deviations due to incomplete client participation. By doing so, the server steers the learning by a small number of representative data, while the clients assist the learning by federation to leverage the huge amount of privately decentralized data ($n_S\gg n_T$).

For SA-FL, we consider the same incomplete client participation setting that induces a dataset $S_{\mathcal{D}}\sim\mathcal{D}$ with cardinality n_S and $\mathcal{D}\neq P$ (i.e., Q2). As a result, the learning process is to minimize $\mathcal{R}_P(h)$ by utilizing $(\mathcal{X},\mathcal{Y})^{n_T+n_S}$ to learn a hypothesis $h\in\mathcal{H}$. For notional clarity, we assume the joint dataset $S_Q=(S_{\mathcal{D}}\cup T)\sim Q$ with cardinality n_T+n_S for some distribution Q. Before deriving the generalization error bound for SA-FL, we state the following assumptions and definitions.

Assumption 1 (Noise Condition)**.** *Suppose h_P^* and h_Q^* exist. There exist $\beta_P,\beta_Q\in[0,1]$ and $\alpha_P,\alpha_Q>0$ s.t., $\mathbb{P}_{x\sim P}(h(x)\neq h_P^*(x))\leq\alpha_P[\varepsilon_P(h)]^{\beta_P}$, $\mathbb{P}_{x\sim Q}(h(x)\neq h_Q^*(x))\leq\alpha_q[\varepsilon_Q(h)]^{\beta_Q}$.*

This assumption is a traditional noise model known as the Bernstein class condition, which has been widely used in the literature (Massart & Nédélec, 2006; Koltchinskii, 2006; Hanneke, 2016).

Assumption 2 ((α,β)-Positively-Related)**.** *Distributions P and Q are said to be (α,β)-positively-related if there exist constants $\alpha\geq0$ and $\beta\geq0$ such that $|\varepsilon_P(h)-\varepsilon_Q(h)|\leq\alpha[\varepsilon_Q(h)]^{\beta},\forall h\in\mathcal{H}$.*

Assumption 2 specifies a stronger constraint between distributions P and Q. It implies that the difference of excess error for one hypothesis $h\in\mathcal{H}$ between P and Q is bounded by the excess error of Q in some exponential form. Assumption 2 is one of the major *novelty* in our paper and unseen in the literature. We note that this (α,β)-positively-related condition is a mild condition. To see this, consider the following "one-dimensional" example for simplicity. Let $\mathcal{H}$ be the class of hypotheses defined on the real line: $\{h_t=t,t\in R\}$, and let two uniform distributions be $P:=\mathcal{U}[a,b]$ and $Q:=\mathcal{U}[a',b']$. Due to the incomplete client sampling in FL, the support of Q is a subset of that of P, i.e., $a\leq a'\leq b'\leq b$. Denote the target hypothesis $t^*\in[a',b']$. Then, for any hypothesis h_t with threshold t, we have $\epsilon_P(h_t)=\frac{|t-t^*|}{b-a}$ and $\epsilon_Q(h_t)=\frac{|t-t^*|}{b'-a'}$. That is, our "$(\alpha,\beta)$-Positively-Related" holds for $\alpha=1-\frac{b'-a'}{b-a}$

and $\beta = 1$. The above "one-dimensional" example can be further extended to general high-dimensional cases as follows. Intuitively, the difference of excess errors of P and Q (i.e., $|\epsilon_P(h) - \epsilon_Q(h)|$) is a function in the form of $\int_S |Q_X - P_X| dS$ for a common support domain $S \subset supp(Q)$. Thus, the "(α, β)-Positively-Related" condition can be written as $|\int_S Q_X dS - \int_S P_X dS| \leq \alpha(\int_S Q_X)^\beta$. If distribution Q has more probability mass over S than distribution P, choosing $\beta = 1$ and α to be a sufficiently large constant clearly satisfies the (α, β)-positively-related condition. Otherwise, letting $\beta \to 0$ and choosing α to be a sufficiently large constant satisfies the (α, β)-positively-related condition with probability one.

With the above assumption and definition, we have the following generation error bound for SA-FL, which shows that SA-FL is PAC-learnable:

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 2, then with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, it holds that*

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

where $d_{\mathcal{H}}$ is the finite VC dimension for hypotheses class $\mathcal{H}$.

Note the generalization error bound of centralized learning is $\widetilde{\mathcal{O}}((\frac{d_{\mathcal{H}}}{n})^{\frac{1}{2-\beta_Q}})$ (hiding logarithmic factors) with n samples in total and noise parameter β_Q (Hanneke, 2016). Note that when $\beta \geq 1$, the first term in Eq. (1) dominates. Hence, Theorem 2 implies that the generalization error bound in this case for SA-FL *matches* that of centralized learning (with dataset size $n_T + n_S$). Meanwhile, for $0 < \beta < 1$, compared with solely training on server's dataset T, SA-FL exhibits an improvement from $\widetilde{\mathcal{O}}((\frac{1}{n_T})^{\frac{1}{2-\beta_Q}})$ to $\widetilde{\mathcal{O}}((\frac{1}{n_T+n_S})^{\frac{\beta}{2-\beta_Q}})$.

Note that SA-FL shares some similarity with the domain adaptation problem, where the learning is on Q but the results will be adapted to P. In what follows, we offer some deeper insights between the two by answering two key questions: *1) What is the difference between SA-FL and domain adaptation (or transfer learning)?* and *2) Why is SA-FL from Q to P PAC-learnable, but FL from D to P with incomplete client participation not PAC-learnable (as indicated in Theorem 1)?*

To answer these questions, we illustrate the distribution relationships for domain adaptation and federated learning, in Fig. 1, respectively. In domain adaptation, the target P and source Q distributions often have overlapping support but there also exists *distinguishable difference*. In contrast, the two distributions P and Q in SA-FL happen to share exactly the *same support* with different density, since Q is a *mixture* of D and P. As a result, the known bounds in domain adaptation (or transfer learning) are pessimistic for SA-FL. For example, the $dist(P, Q)$ in $d_{\mathcal{A}}$-divergence and $\mathcal{Y}$-divergence both have non-negligible gaps when applied to SA-FL. Here in Theorem 2, we provide a generalization error bound in terms of the total sample size $n_T + n_S$, thus showing the benefit of SA-FL.

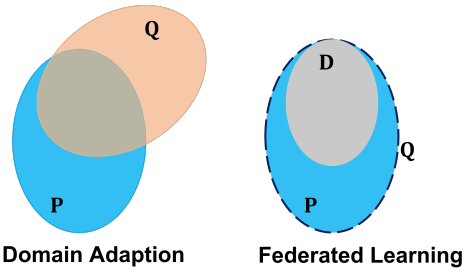

Figure 1. Diagram of distribution supports for domain adaptation and federated learning.

Moreover, for SA-FL, only the auxiliary dataset $T \overset{i.i.d.}{\sim} P$ is directly available to the server. The clients' datasets could be used in SA-FL training, but they are not directly accessible due to privacy constraints. Thus, previous methods in domain adaptation (e.g., importance weights-based methods in covariate shift adaptation (Sugiyama et al., 2007a;b)) are *not* applicable since they require the knowledge of density ratio between training and test datasets.

The key difference between FL and SA-FL lies in relations among D, P and Q. For FL, the distance between D and P with incomplete participation could be large due to system and data heterogeneity in the worst-case. More specifically, the support of D could be narrow enough to miss some part of P, resulting in non-vanishing error as indicated in Theorem 1. For SA-FL, distribution Q is a mixture of P and D ($Q = \lambda_1 D + \lambda_2 P$, with $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$), thus having the same support with P. Hence, under Assumption 2, the PAC-learnability is guaranteed. Although we provide a promising bound to show the PAC-learnability of SA-FL in Theorem 2, the superiority of SA-FL over training solely with dataset T in server (i.e., $\widetilde{\mathcal{O}}((\frac{1}{n_T})^{\frac{1}{2-\beta_P}})$) is not always guaranteed as $\beta \to 0$ (i.e., Q becomes increasingly different from P). In what follows, we reveal under what conditions could SA-FL perform *no worse than* centralized learning.

Theorem 3 (Conditions of 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 $\varepsilon_P(\hat{h}_Q^*) = \widetilde{\mathcal{O}}\left((d_{\mathcal{H}}/n_T)^{\frac{1}{2-\beta_P}}\right)$.*

Here, we remark that $\varepsilon_P(h_Q^*) = \mathcal{O}(\mathcal{A}(n_T, \delta))$ is a weaker condition than the $\varepsilon_P(h_Q^*) = 0$ condition and the covariate shift assumption ($P_{Y|X} = Q_{Y|X}$) used in the transfer learn-

ing literatures (Hanneke & Kpotufe, 2019; 2020). Together with the condition $\hat{\mathcal{R}}_P(\hat{h}^*_Q) \leq \hat{\mathcal{R}}_P(h^*_Q)$, the following intermediate result holds: $\hat{\mathcal{R}}_P(\hat{h}^*_Q) - \hat{\mathcal{R}}_P(h^*_P) = \mathcal{O}(A(n_T, \delta))$ (see Lemma 2 in the supplementary material). Intuitively, this states that "if P and Q share enough similarity, then the difference of excess empirical error between $\hat{h}^*_Q$ and h^*_P on P can be bounded." Thus, the excess error of $\hat{h}^*_Q$ shares the same upper bound as that of $\hat{h}^*_P$ in centralized learning. Therefore, Theorem 3 implies that, under mild conditions, SA-FL guarantees the same generalization error upper bound as that of centralized learning, hence being "no worse than" centralized learning with dataset T.

Remark 1. Note that the assumption of having a server-side dataset is not restrictive due to the following reasons. First, such datasets are already available in many FL systems: although not always necessary for training, an auxiliary dataset is often needed for defining FL tasks (e.g., simulation prototyping) before training and model checking after training (e.g., quality evaluation and sanity checking) (McMahan et al., 2021; Wang et al., 2021a). Also, obtaining an auxiliary dataset is affordable since the number of data points required is relatively small, and hence the cost is low. Then, SA-FL can be easily achieved or even with manually labelled data thanks to its small size. It is also worth noting that many works have used such auxiliary datasets in FL for security (Cao et al., 2021), incentive design (Wang et al., 2019), and knowledge distillation (Cho et al., 2021).

Remark 2. It is also worth pointing out that, for ease of illustration, Theorem 2–3 are based on the assumption that the auxiliary dataset $T \overset{i.i.d.}{\sim} P$. Nonetheless, it is of practical importance to consider the scenario where T is sampled from a related but slightly different distribution P' rather than the target distribution P itself. In fact, the above assumption could be relaxed to $T \overset{i.i.d.}{\sim} P'$ for any P' as long as the mixture distribution $Q = \lambda_1 D + \lambda_2 P'$ is (α, β)-positively-related with P. Under such condition, we can show that the main results in Theorem 2–3 still hold.

4. The SAFARI Algorithm for SA-FL

In Section 3, we have shown that SA-FL is PAC-learnable with incomplete client participation. In this section, we turn our attention to the *training* of the SA-FL regime with incomplete client participation (i.e., Q3), which is also underexplored in the literature. First, we note that the standard FedAvg algorithm may fail to converge to a stationary point with incomplete client participation as indicated by previous works (Yang et al., 2022). Now with SA-FL, we aim to answer the following questions:

1) Under SA-FL, how should we appropriately use the server-side dataset to develop training algorithms in the SA-FL regime with provable convergence guarantees?

Algorithm 1 The SAFARI Algorithm for SA-FL.

1: Initialize model $\mathbf{x}_0$, iteration index $t = 0$.
2: **for** $r = 0, \cdots, R-1$ **do**
3: With probability q: ★ client update round $r \in \mathcal{T}_c$
4: The server samples clients S_r.
5: Each client $i \in S_r$ computes in parallel:
6: $\mathbf{x}^i_{r,k+1} = \mathbf{x}^i_{r,k} - \eta_c \nabla F_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k}), k \in [K]$,
7: starting from $\mathbf{x}^i_{r,0} = \mathbf{x}_r$.
8: Send $\mathbf{x}^i_r = \mathbf{x}^i_{r,K+1}$ to server.
9: Server updates: $\mathbf{x}_{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \mathbf{x}^i_r$.
10: Otherwise ★ server update round $r \in \mathcal{T}_s$
11: Server updates: $\mathbf{x}_{r+1} = \mathbf{x}_r - \eta_s \nabla F(\mathbf{x}_r, \xi_r)$.
12: **end for**

2) If Question 1) can be resolved, could we further achieve the same convergence rate in SA-FL training with incomplete client participation as that in traditional FL with ideal client participation?

In this section, we resolve the above questions affirmatively by proposing a new algorithm called SAFARI (server-assisted federated averaging) for SA-FL with theoretically provable convergence guarantees. As shown in Algorithm 1, SAFARI contains two options in each round, client update option or global server update option. For a communication round $r \in \{0, \cdots, R-1\}$, with probability $q \in [0, 1]$, the client update option is chosen (i.e., $r \in \mathcal{T}_c$), where local updates are executed by clients in the current participating client set S_r in a similar fashion as the FedAvg (McMahan et al., 2017). Specifically, the client update option performs the following three steps: 1) Server samples a subset of clients S_r as in conventional FL and synchronizes the latest global model $\mathbf{x}_r$ with each participating clients in S_r (Line 4); 2) All participating clients initialize their local models as $\mathbf{x}_r$ and then perform K local steps following the stochastic gradient descent (SGD) method. Then, each participating client sends its locally updated model $\mathbf{x}^i_r = \mathbf{x}^i_{r,K+1}$ back to the server (Lines 5-8); 3) Upon receiving the local update $\mathbf{x}^i_r$, the server aggregates and updates the global model (Line 9). On the other hand, with probability $1 - q$, the server update option is chosen (i.e,. $r \in \mathcal{T}_s$), where the server updates the global model with its auxiliary data following the SGD (Line 11).

We note that SAFARI can be viewed as a mixture of the FedAvg algorithm with client-side datasets (cf. the client update option) and a centralized SGD algorithm using the server-side dataset only (cf. the server update option), which are governed by a probability parameter q. The basic idea of this two-option approach is to leverage client-side parallel computing to accelerate the training process, while using the server-side dataset to mitigate the bias caused by incomplete client participation. We will show later that, by appropriately choosing the q-value, SAFARI simultane-

ously achieves the stationary point convergence and linear convergence speedup. Before presenting the convergence performance results, we first state three commonly used assumptions in FL.

Assumption 3. *(L-Lipschitz Continuous Gradient) There exists a constant $L > 0$, such that $\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall i \in [M], \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Assumption 4. *(Unbiased Stochastic Gradients with Bounded Variance) The stochastic gradient calculated by the client or server is unbiased with bounded variance: for server, $\mathbb{E}[\nabla F(\mathbf{x}, \xi)] = \nabla F(\mathbf{x})$ and $\mathbb{E}[\|\nabla F(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|^2] \leq \sigma_s^2$; for each client $i \in [M]$, $\mathbb{E}[\nabla F_i(\mathbf{x}, \xi)] = \nabla F_i(\mathbf{x})$, and $\mathbb{E}[\|\nabla F_i(\mathbf{x}, \xi) - \nabla F_i(\mathbf{x})\|^2] \leq \sigma^2$.*

Assumption 5. *(Bounded Gradient Dissimilarity) $\|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \sigma_G^2, \forall i \in [M]$.*

With the assumptions above, we state the main convergence result of SAFARI for non-convex functions as follows:

Theorem 4 (Convergence Rate for SAFARI in Non-Convex Functions)**.** *Under Assumptions 3 - 5, if $\eta_c \leq \frac{1}{4\sqrt{30}LK}$, $\eta_c = \frac{2\eta_s}{K}$, and $q \leq 1/\left(\frac{4\sigma_G^2 - 4G_2(\frac{1}{2K^2} - \frac{2L\eta_s^2}{K^2})}{(1-L\eta_s)G_1} + 1\right)$, then, the sequence $\{\mathbf{x}_r\}$ generated by* SAFARI *satisfies:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla F(\mathbf{x}_r)\|^2 \leq \frac{2(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{R\eta_s} + \delta,$$

where $\delta = L\eta_s(1-q)\sigma_s^2 + \frac{80qL^2\eta_s^2}{K}(\sigma^2 + 6K\sigma_G^2) + \frac{8Lq\eta_s}{mK}\sigma^2$, $G_1 = \max_{r\in\mathcal{T}_s}\|\nabla F(\mathbf{x}_r)\|^2$, and $G_2 = \max_{r\in\mathcal{T}_c}\left\|\frac{1}{m}\sum_{i\in[m]}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2$.

Remark 3. Theorem 4 says that, by using the server-side update with an appropriately chosen q-value, SAFARI effectively mitigates the bias that arises from incomplete client participation. With proper probability q, SAFARI guarantees stationary point convergence in non-convex functions.

By choosing parameters q and the learning rate η appropriately, Theorem 4 immediately implies the following rate:

Corollary 1. *If $\eta_s = \frac{1}{\sqrt{R}}$,* SAFARI *achieves an $\mathcal{O}(\frac{1}{\sqrt{R}})$ convergence rate to a stationary point. If we can further assume the stochastic gradient noise at server's side $\sigma_s^2 = \mathcal{O}(\frac{1}{mK}\sigma^2)$ or $q = \Omega(1 - \frac{1}{mK})$,* SAFARI *achieves an $\mathcal{O}(\frac{1}{\sqrt{mKR}})$ convergence rate to a stationary point, implying a linear speedup of convergence in terms of m and K.*

For strongly convex functions, we have the following convergence results for SAFARI :

Theorem 5 (Convergence Rate for SAFARI in Strongly Convex Functions)**.** *Under Assumptions 3 - 5 and assume each function F_i is $\mu-$strongly convex, if $\eta_s \leq \frac{2}{L+\mu}$ and $q \leq 1/\left(1 + \frac{\frac{4\bar{\eta}}{\mu^2}\left(1+\frac{30L\bar{\eta}}{\mu}(1+\frac{2L\bar{\eta}}{\mu})\right)G_3 - \frac{4}{\mu}G_4}{\left(\frac{1}{L+\mu} - \frac{(L+\mu)^2\bar{\eta}}{4L^2\mu^2}\right)G_3}\right)$, then, the sequence $\{\mathbf{x}_r\}$ generated by* SAFARI *satisfies:*

$$\mathbb{E}\|\mathbf{x}_R - \mathbf{x}_*\|^2 \leq (1-\bar{\eta})^R\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \bar{\delta},$$

where $\bar{\delta} = \frac{8q\eta_c K}{\mu}\sigma_G^2 + \frac{2q\eta_c}{\mu m}\sigma^2 + \frac{4qL(1+K\eta_c L)}{\mu} \times \left[5K\eta_c^2(\sigma^2 + 6K\sigma_G^2)\right] + (1-q)\frac{L+\mu}{2L\mu}\eta_s\sigma_s^2$, $G_3 = \|\nabla F(\mathbf{x}_r)\|^2$, $G_4 = \frac{1}{m}\sum_{i\in S_r}[F_i(\mathbf{x}_r) - F_i(\mathbf{x}_*)]$, and $\bar{\eta} = \frac{\eta_c K\mu}{2} = \frac{2\eta_s L\mu}{L+\mu}$.

The following results immediately follow from Theorem 5:

Corollary 2. *If $\eta_c = \Omega(\frac{log(R)}{R})$ and $\eta_s = \Omega(\frac{log(R)}{R})$,* SAFARI *achieves an $\tilde{\mathcal{O}}(\frac{1}{R})$ convergence rate.*

Remark 4. In the strongly convex setting, Theorem 5 shows that, with proper hyperparameters, SAFARI achieves convergence guarantees and can effectively mitigate the impacts of incomplete client participation.

Remark 5. We note that SAFARI is a unifying framework that includes two classic algorithms as special cases under two extreme settings: i) the i.i.d. client-side data case and ii) the heterogeneous client-side data case with unbounded gradient dissimilarity. In the i.i.d. case, the client-side data are homogeneous, i.e., $F_i(\mathbf{x}) = F$ and $\sigma_G = 0$. In this ideal setting, we can simply choose $q = 1$ and SAFARI reduces to the classical FedAvg algorithm. In the heterogeneous case with unbounded gradient dissimilarity (i.e., $\sigma_G \to \infty$), we can set $q = 0$ (i.e., $|\mathcal{T}_c| = 0$) such that SAFARI reduces to the centralized SGD algorithm. In this setting, Theorem 4 and 5 recover the classic SGD bounds by cancelling the σ_G-dependent terms in the bound.

Remark 6. Corollary 1 and 2 suggest that, thanks to the two "control knobs", learning rate η and q in SAFARI , under mild conditions, we can avoid the limitation of conventional FL algorithms. For example, FedAvg with incomplete client participation can only converge to an error ball dependent on the data heterogeneity parameter σ_G (Yang et al., 2022). In SA-FL, SAFARI with incomplete client participation can still achieve the same convergence rates as these of classic FedAvg algorithms with ideal client participation: $\mathcal{O}(1/\sqrt{mKR})$ for non-convex functions (Yang et al., 2021b) and $\tilde{\mathcal{O}}(1/R)$ for strongly convex functions (Li et al., 2020b).

5. Numerical results

In this section, we conduct numerical experiments to verify our theoretical results using 1) logistic regression (LR) on MNIST dataset (LeCun et al., 1998), 2) convolutional neural network (CNN) on CIFAR-10 dataset (Krizhevsky et al., 2009). To simulate data heterogeneity, we distribute

Table 1. Test accuracy (%) for FedAvg.

DATASET	s	NON-I.I.D. INDEX (p)			
		10	5	2	1
MNIST	0	92.69	89.49	86.17	84.49
	2	92.64	89.11	86.54	71.58
	4	92.62	88.81	77.81	57.05
CIFAR-10	0	81.12	79.42	78.22	75.7
	2	79.97	78.54	76.68	64.56
	4	77.78	75.55	67.34	50.7

the data into each client evenly in a label-based partition, following the same process as in previous works (McMahan et al., 2017; Yang et al., 2021b; Li et al., 2020b). As a result, we can use a parameter $p \in \{1, 2, 5, 10\}$ to represent the classes of labels in each client's dataset, which serves as an index of data heterogeneity level (non-i.i.d. index). The smaller p-value, the more heterogeneous the data among clients. To mimic incomplete client participation, we force s clients to be excluded. We can use $s \in \{0, 2, 4\}$ as an index to represent the degree of incomplete client participation. In our experiments, there are $M = 10$ clients in total, and $m = 5$ clients participate in the training in each communication round, who are uniformly sampled from the $M - s$ clients. We use FedAvg without any server-side dataset as the baseline to compare with SAFARI with auxiliary data size $\{50, 100, 500, 1000\}$ for MNIST and $\{500, 5000, 10000\}$ for CIFAR10. So in each dataset, we have at least $4 \times 3 \times 4 = 48$ sets of experiment for ablation study. Due to space limitation, we highlight the key observations in this section, and relegate all other experimental details and results to the supplementary material.

1) Performance Degradation of Incomplete Client Participation: As shown in Table 1, a distinct and non-trivial decline in performance is observed for FedAvg when confronted with incomplete client participation. As the value of s increases, it signifies progressively more incomplete client participation. Upon comparing scenarios with $s = 0$ and $s = 4$, a significant reduction in test accuracy becomes apparent, reaching up to 27.44% for MNIST and 25% for CIFAR10. It is noteworthy that such performance degradation is also contingent on data heterogeneity, denoted by p. In the case of IID data ($p = 10$), only a negligible decrease in test accuracy is observed. For instance, in MNIST, the accuracy slightly drops from 92.69% for $s = 0$ to 92.62% for $s = 4$. In CIFAR10, the accuracy decreases from 81.12% for $s = 0$ to 77.78% for $s = 4$. These results empirically validate the worst-case analysis in Theorem 1 and serve as the primary motivation for the development of SA-FL.

2) Improvement of the SAFARI Algorithm under Incomplete Client Participation: The improvements of our SAFARI can be observed in two aspects: improved test accuracy and faster convergence rate.

I. Improved test accuracy. In Table 2 and 3, we show the test accuracy improvement of our SAFARI algorithm compared with that of FedAvg in standard FL. In Table 2, we can observe that even with a few server participation with a probability of 0.2, there is a non-negligible improvement in test accuracy. In Table 3, the key observation is that, with a *small amount* of auxiliary data at the server, there is a significant increase of test accuracy for our SAFARI algorithm. For example, with only 50 data samples at the server (0.1% of the total training data), there is a 16.65% test accuracy increase. With 1000 data samples, the improvement reaches 31.07%. This verifies the effectiveness of our SA-FL framework and our SAFARI algorithm. Another observation is that for nearly homogeneous case (e.g., from $p = 10$ to $p = 5$), there is no statistical difference with or without auxiliary data at the server (denoted by '-' in Table 3. This is consistent with the previous observations of negligible degradation in cases with homogeneous data across clients.

Table 2. Test accuracy improvement (%) for SAFARI ($q = 0.8$) compared with FedAvg. '-' means "no statistical difference within 2% error bar".

DATASET	s	NON-I.I.D. INDEX (p)			
		10	5	2	1
MNIST	0	-	-	3.13	4.47
	2	-	-	2.01	16.53
	4	-	-	10.69	31.07
CIFAR-10	0	-	-	-	2.57
	2	-	-	-	12.57
	4	-	2.93	9.32	23.86

Table 3. Test accuracy improvement (%) for SAFARI compared with FedAvg on MNIST ($q = 0.8$, $s = 4$). '-' means "no statistical difference within 2% error bar".

DATASET SIZE	NON-I.I.D. INDEX (p)			
	10	5	2	1
50	-	-	4.82	16.65
100	-	-	6.87	20.26
500	-	-	9.16	29.82
1000	-	-	10.69	31.07

II. Faster convergence rate. In Fig. 2, we show the convergence processes of SAFARI on CIFAR-10 under incomplete client participation ($s = 4$) with non-i.i.d. data ($p = 1$). We can see clearly that the convergence of SAFARI is accelerating and the test accuracy increases as more data are employed at the server.

6. Conclusion

In this paper, we rigorously investigated the server-assisted federated learning (SA-FL) framework (i.e., to deploy an auxiliary dataset at the server), which has been increasingly adopted in practice to mitigate the impacts of incomplete client participation in conventional FL. To characterize the benefits of SA-FL, we first showed that conventional FL is *not* PAC-learnable under incomplete client participation by establishing a fundamental generalization error lower

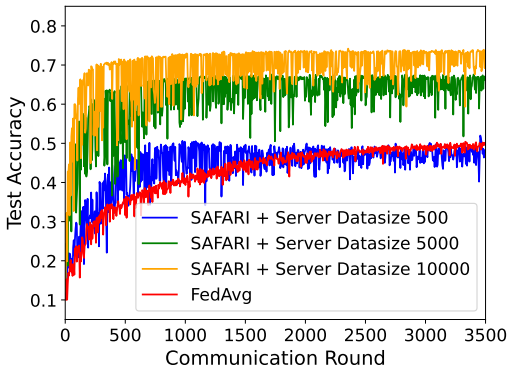

Figure 2. Comparison of test accuracy on CIFAR-10 ($s = 4$, $p = 1$, $q = 0.4$).

bound. Then, we showed that SA-FL is able to revive the PAC-learnability of conventional FL under incomplete client participation. Upon resolving the PAC-learnability challenge, we proposed a new SAFARI (server-assisted federated averaging) algorithm that enjoys convergence guarantee and the same level of communication efficiency as that of conventional FL. Extensive numerical results also validated our theoretical findings.

Acknowledgements

This work is supported in part by AI Seed Funding and the GWBC Award at RIT, as well as NSF grants CAREER CNS-2110259, CNS-2112471, and IIS-2324052. The authors also thank Mr. Zhe Li for his assistance with the part of the experiments.

Impact Statement

This paper delves into the realm of federated learning, a cutting-edge paradigm in machine learning that leverages decentralized training across multiple devices while preserving data privacy. The primary objective of our research is to contribute to the advancement of federated learning methodologies, addressing challenges and exploring opportunities for enhancing model performance in a privacy-preserving manner. As with any technological innovation, our work on federated learning has many potential societal consequences, none which we feel must be specifically highlighted here.

References

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Avdiukhin, D. and Kasiviswanathan, S. Federated learning under arbitrary communication patterns. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 425–435. PMLR, 18–24 Jul 2021.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecnỳ, J., Mazzocchi, S., McMahan, H. B., et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

Bshouty, N. H., Eiron, N., and Kushilevitz, E. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2): 255–275, 2002.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Cao, X., Fang, M., Liu, J., and Gong, N. Z. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *ISOC Network and Distributed Systems Security (NDSS) Symposium*, abs/2012.13995, 2021.

Chen, W., Horvath, S., and Richtarik, P. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

Cho, Y. J., Wang, J., Chiruvolu, T., and Joshi, G. Personalized federated learning for heterogeneous clients with clustered knowledge transfer. *arXiv preprint arXiv:2109.08119*, 2021.

Cho, Y. J., Sharma, P., Joshi, G., Xu, Z., Kale, S., and Zhang, T. On the convergence of federated averaging with cyclic client participation. *arXiv preprint arXiv:2302.03109*, 2023.

Condat, L., Agarskỳ, I., Malinovsky, G., and Richtárik, P. Tamuna: Doubly accelerated federated learning with local training, compression, and partial participation. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023.

David, S. B., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, 2010.

Grudzień, M., Malinovsky, G., and Richtárik, P. Can 5th generation local training methods support client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pp. 1055–1092. PMLR, 2023.

Gu, X., Huang, K., Zhang, J., and Huang, L. Fast federated learning in the presence of arbitrary device unavailability. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Hanneke, S. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17 (1):4667–4721, 2016.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning. In *NeurIPS*, 2019.

Hanneke, S. and Kpotufe, S. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Breaking the centralized barrier for cross-device federated learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Khanduri, P., SHARMA, P., Yang, H., Hong, M., Liu, J., Rajawat, K., and Varshney, P. STEM: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Koloskova, A., Stich, S. U., and Jaggi, M. Sharper convergence guarantees for asynchronous sgd for distributed and federated learning. *Advances in Neural Information Processing Systems*, 35:17202–17215, 2022.

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Konstantinov, N., Frantar, E., Alistarh, D., and Lampert, C. On the sample complexity of adversarial multi-source pac learning. In *International Conference on Machine Learning*, pp. 5416–5425. PMLR, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In Dhillon, I., Papailiopoulos, D., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020a.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020b.

Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., and Feng, J. No fear of heterogeneity: Classifier calibration for federated learning with non-IID data. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Malinovsky, G., Horváth, S., Burlachenko, K., and Richtárik, P. Federated learning with regularized client participation. *arXiv preprint arXiv:2302.03662*, 2023.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, H. B. et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1), 2021.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pp. 124–138. Springer, 2012.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Murata, T. and Suzuki, T. Bias-variance reduced local sgd for less heterogeneous federated learning. *arXiv preprint arXiv:2102.03198*, 2021.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konecný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Ruan, Y., Zhang, X., Liang, S.-C., and Joe-Wong, C. Towards flexible device participation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3403–3411. PMLR, 2021.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007a.

Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pp. 1433–1440. Citeseer, 2007b.

Wang, G., Dang, C. X., and Zhou, Z. Measure contribution of participants in federated learning. *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2597–2604, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021a.

Wang, L., Xu, S., Wang, X., and Zhu, Q. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10165–10173, 2021b.

Wang, S. and Ji, M. A unified analysis of federated learning with arbitrary client participation. *arXiv preprint arXiv:2205.13648*, 2022.

Wang, S. and Ji, M. A lightweight method for tackling unknown participation probabilities in federated averaging. *arXiv preprint arXiv:2306.03401*, 2023.

Xie, C., Koyejo, S., and Gupta, I. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

Yang, C., Wang, Q., Xu, M., Chen, Z., Bian, K., Liu, Y., and Liu, X. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *Proceedings of the Web Conference 2021*, pp. 935–946, 2021a.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021b.

Yang, H., Zhang, X., Khanduri, P., and Liu, J. Anarchic federated learning. In *International Conference on Machine Learning*, pp. 25331–25363. PMLR, 2022.

Zhang, X., Hong, M., Dhople, S., Yin, W., and Liu, Y. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

A. Proofs

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

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 $\epsilon = \frac{1-\omega}{8}$.

First we show that the rare points x_2 appears at most $(1-\omega)Mn$ 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,

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

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

Next, we consider the following sampling process with dataset $S = \{(x_1^{'}, f(x_1^{'})), \ldots, (x_{M\times n}^{'}, f(x_{M\times n}^{'}))\}$: 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 $\hat{s} < (1-\omega)MN$.

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:

$$\begin{aligned}
\mathcal{R}(h_s, f_1) + \mathcal{R}(h_s, f_2) &= \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_1(x) \cap x \neq x_1] + \mathbb{P}_P[\mathcal{L}(\mathcal{F}(S))(x) \neq f_2(x) \cap x \neq x_1] && (2)\\
&= \mathbf{1}_{\mathcal{L}(\mathcal{F}(S))(x_1)\neq f_1(x_1)}\mathbb{P}(x_2) + \mathbf{1}_{\mathcal{L}(\mathcal{F}(S))(x_1)\neq f_2(x_2)}\mathbb{P}(x_1) && (3)\\
&= 4\epsilon. && (4)
\end{aligned}$$

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:

$$\begin{aligned}
2\epsilon \leq \mathbb{E}_{S\in\mathcal{S}}\mathcal{R}(h_s, f) &= \sum_{S:\mathcal{R}(h_s,f)\geq\epsilon} \mathcal{R}(h_s, f)\mathbb{P}[\mathcal{R}(h_s, f)] + \sum_{S:\mathcal{R}(h_s,f)<\epsilon} \mathcal{R}(\mathcal{R}(h_s, f)\mathbb{P}[\mathcal{R}(h_s, f)] && (5)\\
&\leq 4\epsilon\mathbb{P}_{S\in\mathcal{S}}[\mathcal{R}(h_s, f) \geq 4\epsilon] + \epsilon(1 - \mathbb{P}_{S\in\mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon]) && (6)\\
&= \epsilon + 3\epsilon\mathbb{P}_{S\in\mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon]. && (7)
\end{aligned}$$

That is,

$$\mathbb{P}_{S\in\mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon] \geq \frac{1}{3}. \tag{8}$$

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:

$$\begin{aligned}
\mathbb{P}_{S\sim P}[\mathcal{R}_P(\mathcal{L}(\mathcal{F}(S)), f) \geq \epsilon] &\geq \mathbb{P}_{S\sim P}[\mathcal{R}(h_s, f) \geq \epsilon] && (9)\\
&\geq \mathbb{P}_{S\in\mathcal{S}}[\mathcal{R}(h_s, f) \geq \epsilon]\mathbb{P}[S \in \mathcal{S}] && (10)\\
&> \frac{1}{3}\frac{3}{20} = \frac{1}{20}. && (11)
\end{aligned}$$

□

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 2, then with probability at least $1-\delta$ for any $\delta \in (0, 1)$, it holds that*

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

where $d_{\mathcal{H}}$ is the finite VC dimension for hypotheses class $\mathcal{H}$.

Proof.

$$\varepsilon_P(\hat{h}_Q^*) = \mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) \tag{12}$$
$$= [\mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) - (\mathcal{R}_Q(\hat{h}_Q^*) - \mathcal{R}_Q(h_Q^*))] + \mathcal{R}_Q(\hat{h}_Q^*) - \mathcal{R}_Q(h_Q^*) \tag{13}$$
$$\leq |\varepsilon_P(\hat{h}_Q^*) - \varepsilon_Q(\hat{h}_Q^*)| + \varepsilon_Q(\hat{h}_Q^*) \tag{14}$$
$$\leq \alpha\varepsilon_Q(\hat{h}_Q^*)^\beta + \varepsilon_Q(\hat{h}_Q^*). \tag{15}$$

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 $1-\delta$, $\forall h, \hat{h} \in \mathcal{H}$,*

$$\mathcal{R}(h) - \mathcal{R}(\hat{h}) \leq \hat{\mathcal{R}}(h) - \hat{\mathcal{R}}(\hat{h}) + c\sqrt{\min\{\mathbb{P}_S(h \neq \hat{h}), \hat{\mathbb{P}}_S(h \neq \hat{h})\}A(m,\delta)} + cA(m,\delta), \tag{16}$$
$$\frac{1}{2}\mathbb{P}_S(h \neq \hat{h}) - cA(m,\delta) \leq \hat{\mathbb{P}}_S(h \neq \hat{h}) \leq 2\mathbb{P}_S(h \neq \hat{h}) + cA(m,\delta), \tag{17}$$
$$\varepsilon_Q(\hat{h}_Q^*) = [A(m,\delta)]^{\frac{1}{2-\beta_Q}}, \tag{18}$$

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

□

Theorem 3 (Conditions of 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 $\varepsilon_P(\hat{h}_Q^*) = \widetilde{\mathcal{O}}\left((d_\mathcal{H}/n_T)^{\frac{1}{2-\beta_P}}\right)$.*

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.

$$\varepsilon_P(\hat{h}_Q^*) = \mathcal{R}_P(\hat{h}_Q^*) - \mathcal{R}_P(h_P^*) \tag{19}$$
$$\leq \hat{\mathcal{R}}_P(\hat{h}_Q^*) - \hat{\mathcal{R}}_P(h_P^*) + c\sqrt{\min\{P(\hat{h}_Q^* \neq h_P^*), \hat{P}(\hat{h}_Q^* \neq h_P^*)\}A(n_T,\delta)} + cA(n_T,\delta) \tag{20}$$
$$\leq c\sqrt{\varepsilon_P^{\beta_P}(\hat{h}_Q^*)A(n_T,\delta)} + cA(n_T,\delta). \tag{21}$$

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.

$$\hat{\mathcal{R}}_P(\hat{h}_Q^*) - \hat{\mathcal{R}}_P(h_P^*) \leq \hat{\mathcal{R}}_P(h_Q^*) - \hat{\mathcal{R}}_P(h_P^*) \tag{22}$$

$$\leq \mathcal{R}_P(h_Q^*) - \mathcal{R}_P(h_P^*) + c\sqrt{\min\{P(h_Q^* \neq h_P^*), \hat{P}(h_Q^* \neq h_P^*)\}A(n_T,\delta)} + cA(n_T,\delta) \tag{23}$$

$$= \varepsilon_P(h_Q^*) + \mathcal{O}(A(n_T,\delta)). \tag{24}$$

□

Theorem 4 (Convergence Rate for SAFARI in Non-Convex Functions)**.** *Under Assumptions 3 - 5, if* $\eta_c \leq \frac{1}{4\sqrt{30}LK}$, $\eta_c = \frac{2\eta_s}{K}$, *and* $q \leq 1/\left(\frac{4\sigma_G^2 - 4G_2(\frac{1}{2K^2} - \frac{2L\eta_s^2}{K^2})}{(1-L\eta_s)G_1} + 1\right)$, *then, the sequence* $\{\mathbf{x}_r\}$ *generated by* SAFARI *satisfies:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla F(\mathbf{x}_r)\|^2 \leq \frac{2(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{R\eta_s} + \delta,$$

where $\delta = L\eta_s(1-q)\sigma_s^2 + \frac{80qL^2\eta_s^2}{K}(\sigma^2 + 6K\sigma_G^2) + \frac{8Lq\eta_s}{mK}\sigma^2$, $G_1 = \max_{r\in\mathcal{T}_s}\|\nabla F(\mathbf{x}_r)\|^2$, *and* $G_2 = \max_{r\in\mathcal{T}_c}\left\|\frac{1}{m}\sum_{i\in[m]}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2$.

Proof. In expectation, we define that there are totally $R_s = |\mathcal{T}_s| = (1-p)R$ rounds for server update, $R_c = |\mathcal{T}_c| = pR$ rounds for client update, and $R = R_s + R_c$,

Taking expectation on the random data samples conditioned on $\mathbf{x}_r$, we can have the following one-step descent when server updates:

$$\mathbb{E}_r[F(\mathbf{x}_{r+1})] \leq F(\mathbf{x}_r) + \left\langle \nabla F(\mathbf{x}_r), \mathbb{E}_r[\mathbf{x}_{r+1} - \mathbf{x}_r]\right\rangle + \frac{L}{2}\mathbb{E}_r[\|\mathbf{x}_{r+1} - \mathbf{x}_r\|^2] \tag{25}$$

$$= F(\mathbf{x}_r) + \left\langle \nabla F(\mathbf{x}_r), \eta_s\mathbb{E}_r[\nabla F(\mathbf{x}_r, \xi_r)]\right\rangle + \frac{L}{2}\eta_s^2\mathbb{E}_r[\|\nabla F(\mathbf{x}_r, \xi_r)\|^2] \tag{26}$$

$$= F(\mathbf{x}_r) - \eta_s\|\nabla F(\mathbf{x}_r)\|^2 + \frac{L\eta_s^2}{2}\|\nabla F(\mathbf{x}_r)\|^2 + \frac{L\eta_s^2}{2}\sigma_s^2. \tag{27}$$

That is,

$$\|\nabla F(\mathbf{x}_r)\|^2 \leq \frac{2}{\eta_s}(F(\mathbf{x}_r) - \mathbb{E}_r[F(\mathbf{x}_{r+1})]) + (L\eta_s - 1)\|\nabla F(\mathbf{x}_r)\|^2 + L\eta_s\sigma_s^2. \tag{28}$$

Similarly, when clients update, we assume there are totally m clients participating in one round, denoted as $[m]$. Then we have:

$$\mathbb{E}_r[F(\mathbf{x}_{r+1})] \leq F(\mathbf{x}_r) + \left\langle \nabla F(\mathbf{x}_r), \mathbb{E}_r[\mathbf{x}_{r+1} - \mathbf{x}_r]\right\rangle + \frac{L}{2}\mathbb{E}_r[\|\mathbf{x}_{r+1} - \mathbf{x}_r\|^2] \tag{29}$$

$$= F(\mathbf{x}_r) + \underbrace{\left\langle \nabla F(\mathbf{x}_r), -\eta_c\mathbb{E}_r[\Delta_r]\right\rangle}_{A_1} + \underbrace{\frac{L}{2}\eta_c^2\mathbb{E}_r[\|\Delta_r\|^2]}_{A_2}. \tag{30}$$

$$A_1 = \left\langle \nabla F(\mathbf{x}_r), -\eta_c\mathbb{E}_r[\Delta_r]\right\rangle \tag{31}$$

$$= \frac{1}{2K}\eta_c\left[-K^2\|\nabla F(\mathbf{x}_r)\|^2 - \|\mathbb{E}_r[\Delta_r]\|^2 + \|K\nabla F(\mathbf{x}_r) - \mathbb{E}_r[\Delta_r]\|^2\right] \tag{32}$$

$$= -\frac{K\eta_c}{2}\|\nabla F(\mathbf{x}_r)\|^2 - \frac{\eta_c}{2K}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{\eta_c}{2K}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\left[\nabla F(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_{r,k}^i)\right]\right\|^2 \tag{33}$$

$$\leq -\frac{K\eta_c}{2}\|\nabla F(\mathbf{x}_r)\|^2 - \frac{\eta_c}{2K}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{\eta_c}{2m}\sum_{i\in S_r}\sum_{k\in[K]}\underbrace{\left\|\nabla F(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2}_{A_3} \tag{34}$$

$$\leq -\frac{K\eta_c}{2}\|\nabla F(\mathbf{x}_r)\|^2 - \frac{\eta_c}{2K}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 \tag{35}$$

$$+ \eta_c K\sigma_G^2 + \eta_c K L^2\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2) + 30K^2\eta_c^2\|\nabla F(\mathbf{x}_r)\|^2\right], \tag{36}$$

where A_3 could be bounded as follows:

$$A_3 = \left\|\nabla F(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 \tag{37}$$

$$= \left\|\nabla F(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_r) + \nabla F_i(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 \tag{38}$$

$$\leq 2\left\|\nabla F(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_r)\right\|^2 + 2\left\|\nabla F_i(\mathbf{x}_r) - \nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 \tag{39}$$

$$\leq 2\sigma_G^2 + 2L^2\left\|\mathbf{x}_r - \mathbf{x}_{r,k}^i\right\|^2 \tag{40}$$

$$\leq 2\sigma_G^2 + 2L^2\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2) + 30K^2\eta_c^2\|\nabla F(\mathbf{x}_r)\|^2\right], \tag{41}$$

where the last inequality follows from the bounded local update step with $\eta_c \leq \frac{1}{8LK}$ (see Lemma 2 in (Yang et al., 2021b) and Lemma 3 in (Reddi et al., 2021)).

$$A_2 = \frac{L}{2}\eta_c^2\mathbb{E}_r[\|\Delta_r\|^2] \tag{42}$$

$$\leq L\eta_c^2\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + L\eta_c^2\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\left[\nabla F_i(\mathbf{x}_{r,k}^i) - \nabla F_i(\mathbf{x}_{r,k}^i,\xi_{r,k}^i)\right]\right\|^2 \tag{43}$$

$$\leq L\eta_c^2\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{L\eta_c^2 K}{m}\sigma^2, \tag{44}$$

where the last inequality is due to the martingale difference sequence $\{\nabla F_i(\mathbf{x}_{r,k}^i) - \nabla F_i(\mathbf{x}_{r,k}^i,\xi_{r,k}^i)\}$ (see Lemma 4 in (Karimireddy et al., 2020)).

Putting pieces together, we have

$$K\eta_c(\frac{1}{2} - 30L^2K^2\eta_c^2)\|\nabla F(\mathbf{x}_r)\|^2 \leq F(\mathbf{x}_r) - \mathbb{E}_r[F(\mathbf{x}_{r+1}) - \frac{\eta_c}{2K}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \eta_c K\sigma_G^2 \tag{45}$$

$$+ \eta_c K L^2\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right] + L\eta_c^2\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{L\eta_c^2K}{m}\sigma^2 \tag{46}$$

If $(\frac{1}{2} - 30L^2K^2\eta_c^2) \geq \frac{1}{4}$ (i.e., $\eta_c \leq \frac{1}{4\sqrt{30}LK}$) and $\eta_c = \frac{2\eta_s}{K}$, we have

$$\|\nabla F(\mathbf{x}_r)\|^2 \leq \frac{4}{K\eta_c}(F(\mathbf{x}_r) - \mathbb{E}_r[F(\mathbf{x}_{r+1})]) + 4\sigma_G^2 \tag{47}$$

$$+ (\frac{4L\eta_c}{K} - \frac{2}{K^2})\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + 20KL^2\eta_c^2(\sigma^2+6K\sigma_G^2) + \frac{4L\eta_c}{m}\sigma^2 \tag{48}$$

$$= \frac{2}{\eta_s}(F(\mathbf{x}_r) - \mathbb{E}_r[F(\mathbf{x}_{r+1})]) + 4\sigma_G^2 \tag{49}$$

$$+ (\frac{8L\eta_s}{K^2} - \frac{2}{K^2})\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{80L^2\eta_s^2}{K}(\sigma^2+6K\sigma_G^2) + \frac{8L\eta_s}{mK}\sigma^2 \tag{50}$$

Note there are totally R_s rounds (T_s as the round indices) for server update and R_c rounds (T_c as the round indices) for client update. Let $R = R_s + R_c$, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla F(\mathbf{x}_r)\|^2 \leq \frac{2}{\eta_s}\frac{1}{R}\sum_{r=1}^{R}(F(\mathbf{x}_r) - F(\mathbf{x}_{r+1})) + \frac{1}{R}\sum_{r\in T_s}(L\eta_s - 1)\|\nabla F(\mathbf{x}_r)\|^2 + \frac{L\eta_s R_s}{R}\sigma_s^2 \tag{51}$$

$$+ \frac{4R_c}{R}\sigma_G^2 + \frac{1}{R}\sum_{r\in T_c}(\frac{8L\eta_s}{K^2} - \frac{2}{K^2})\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 + \frac{80R_cL^2\eta_s^2}{KR}(\sigma^2 + 6K\sigma_G^2) + \frac{8LR_c\eta_s}{mKR}\sigma^2 \tag{52}$$

$$\leq \frac{2(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{R\eta_s} + \frac{L\eta_s R_s}{R}\sigma_s^2 + \frac{80R_cL^2\eta_s^2}{KR}(\sigma^2 + 6K\sigma_G^2) + \frac{8LR_c\eta_s}{mKR}\sigma^2, \tag{53}$$

where the last inequality follows from

$$4R_c\sigma_G^2 \leq \sum_{r\in T_s}(1 - L\eta_s)\|\nabla F(\mathbf{x}_r)\|^2 + \sum_{r\in T_c}(\frac{2}{K^2} - \frac{8L\eta_s}{K^2})\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2 \tag{54}$$

$$\leq R_s(1 - L\eta_s)G_1 + R_c(\frac{2}{K^2} - \frac{8L\eta_s}{K^2})G_2, \tag{55}$$

where $G_1 = \max_{r\in T_s}\|\nabla F(\mathbf{x}_r)\|^2$ and $G_2 = \max_{r\in T_c}\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2$. That is the requirement on q such that $q \leq 1/\left(\frac{4\sigma_G^2 - 4G_2(\frac{1}{2K^2} - \frac{2L\eta_s^2}{K^2})}{(1-L\eta_s)G_1} + 1\right)$.

□

A.1. Discussions

We want to cast caveats on Corollary 1. The results in Corollary 1 does not hold in *arbitrary* cases. Specifically, Corollary 1 requires both $R_s \geq \frac{4\sigma_G^2 - 4G_2(\frac{1}{2K^2} - \frac{2L\eta_s^2}{K^2})}{(1-L\eta_s)G_1}R_c$ and $R_s = \mathcal{O}(\frac{R}{mK})$. In other words, we need $\frac{4\sigma_G^2 - 4G_2(\frac{1}{2K^2} - \frac{2L\eta_s^2}{K^2})}{(1-L\eta_s)G_1} \leq \frac{R_s}{R_c} \leq \frac{c}{mK(1-\frac{c}{mK})}$, where c is a constant. Approximately, $\frac{R_s}{R_c} = constant$. Due to $R_s + R_c = R$, we can see that $R_c = \Omega(R)$. So the convergence rate is $\mathcal{O}(\frac{1}{\sqrt{mKR}})$, which is the same order of $\mathcal{O}(\frac{1}{\sqrt{mKR_c}})$.

Theorem 5 (Convergence Rate for SAFARI in Strongly Convex Functions)**.** *Under Assumptions 3 - 5 and assume each function F_i is $\mu-$strongly convex, if $\eta_s \leq \frac{2}{L+\mu}$ and $q \leq 1/\left(1 + \frac{\frac{4\bar{\eta}}{\mu^2}\left(1+\frac{30L\bar{\eta}}{\mu}(1+\frac{2L\bar{\eta}}{\mu})\right)G_3 - \frac{4}{\mu}G_4}{\left(\frac{1}{L+\mu} - \frac{(L+\mu)^2\bar{\eta}}{4L^2\mu^2}\right)G_3}\right)$, then, the sequence $\{\mathbf{x}_r\}$ generated by* SAFARI *satisfies:*

$$\mathbb{E}\|\mathbf{x}_R - \mathbf{x}_*\|^2 \leq (1 - \bar{\eta})^R\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \bar{\delta},$$

where $\bar{\delta} = \frac{8q\eta_cK}{\mu}\sigma_G^2 + \frac{2q\eta_c}{\mu m}\sigma^2 + \frac{4qL(1+K\eta_cL)}{\mu} \times \left[5K\eta_c^2(\sigma^2 + 6K\sigma_G^2)\right] + (1-q)\frac{L+\mu}{2L\mu}\eta_s\sigma_s^2$, $G_3 = \|\nabla F(\mathbf{x}_r)\|^2$, $G_4 = \frac{1}{m}\sum_{i\in S_r}[F_i(\mathbf{x}_r) - F_i(\mathbf{x}_*)]$, and $\bar{\eta} = \frac{\eta_cK\mu}{2} = \frac{2\eta_sL\mu}{L+\mu}$.

Proof. Taking expectation on the random data samples conditioned on $\mathbf{x}_r$, we can have the following one-step descent as classic stochastic gradient descent method when server updates:

$$\mathbb{E}_r\|\mathbf{x}_{r+1} - \mathbf{x}_*\|^2 = \mathbb{E}_r\|\mathbf{x}_r - \eta_s\nabla F(\mathbf{x}_r, \xi_r) - \mathbf{x}_*\|^2 \tag{56}$$

$$\leq \|\mathbf{x}_r - \mathbf{x}_*\|^2 + \eta_s^2\|\nabla F(\mathbf{x}_r)\|^2 + \eta_s^2\sigma_s^2 - 2\eta_s\langle\mathbf{x}_r - \mathbf{x}_*, \nabla F(\mathbf{x}_r)\rangle \tag{57}$$

$$\leq \left[1 - \frac{2\eta_sL\mu}{L+\mu}\right]\|\mathbf{x}_r - \mathbf{x}_*\|^2 + \eta_s(\eta_s - \frac{2}{L+\mu})\|\nabla F(\mathbf{x}_r)\|^2 + \eta_s^2\sigma_s^2 \tag{58}$$

where the second last inequality is due to the $\mu-$strongly convex property (see Lemma 3.11 in (Bubeck et al., 2015)) and the last inequality follows from the fact $\eta_s \leq \frac{2}{L+\mu}$ and $\mu-$strongly convex property $\|\nabla F(\mathbf{x})\|^2 \geq 2\mu(F(\mathbf{x}) - F(\mathbf{x}_*))$.

When arbitrary clients participate, we have

$$\mathbb{E}_r\|\mathbf{x}_{r+1}-\mathbf{x}_*\|^2 = \mathbb{E}_r\|\mathbf{x}_r-\eta_c\Delta_r-\mathbf{x}_*\|^2 \tag{59}$$

$$\leq \|\mathbf{x}_r-\mathbf{x}_*\|^2+\eta_c^2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i,\xi_{r,k}^i)\right\|^2-2\eta_c\big\langle\mathbf{x}_r-\mathbf{x}_*,\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\big\rangle \tag{60}$$

$$\leq \|\mathbf{x}_r-\mathbf{x}_*\|^2+\eta_c^2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2-2\eta_c\big\langle\mathbf{x}_r-\mathbf{x}_*,\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\big\rangle \tag{61}$$

$$\leq \|\mathbf{x}_r-\mathbf{x}_*\|^2+\eta_c^2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2 \tag{62}$$

$$-2\eta_c\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\Big[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)+\frac{\mu}{4}\|\mathbf{x}_r-\mathbf{x}_*\|^2-L\|\mathbf{x}_r-\mathbf{x}_{r,k}^i\|^2\Big] \tag{63}$$

$$\leq (1-\frac{\eta_cK\mu}{2})\|\mathbf{x}_r-\mathbf{x}_*\|^2+\eta_c^2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2 \tag{64}$$

$$-\frac{2\eta_cK}{m}\sum_{i\in S_r}[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)]+\frac{2\eta_cL}{m}\sum_{i\in S_r}\sum_{k\in[K]}\|\mathbf{x}_r-\mathbf{x}_{r,k}^i\|^2 \tag{65}$$

$$\tag{66}$$

where the second last inequality is due to $\big\langle\nabla f(\mathbf{x}),\mathbf{z}-\mathbf{y}\big\rangle\geq f(\mathbf{z})-f(\mathbf{y})+\frac{\mu}{4}\|\mathbf{z}-\mathbf{y}\|^2-L\|\mathbf{z}-\mathbf{x}\|^2$ for any $\mu-$strongly convex and L-smooth function f (see Lemma 5 in (Karimireddy et al., 2020)).

$$\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)\right\|^2=\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)-\nabla F_i(\mathbf{x}_r)+\nabla F_i(\mathbf{x}_r)\right\|^2 \tag{67}$$

$$\leq 2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_{r,k}^i)-\nabla F_i(\mathbf{x}_r)\right\|^2+2\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_r)\right\|^2 \tag{68}$$

$$\leq \frac{2K}{m}\sum_{i\in S_r}\sum_{k\in[K]}\mathbb{E}_r\left\|\nabla F_i(\mathbf{x}_{r,k}^i)-\nabla F_i(\mathbf{x}_r)\right\|^2+4\mathbb{E}_r\left\|\frac{1}{m}\sum_{i\in S_r}\sum_{k\in[K]}\nabla F_i(\mathbf{x}_r)-\nabla F(\mathbf{x}_r)\right\|^2+4K^2\mathbb{E}_r\left\|\nabla F(\mathbf{x}_r)\right\|^2 \tag{69}$$

$$\leq \frac{2KL^2}{m}\sum_{i\in S_r}\sum_{k\in[K]}\mathbb{E}_r\left\|\mathbf{x}_{r,k}^i-\mathbf{x}_r\right\|^2+4K^2\sigma_G^2+4K^2\mathbb{E}_r\left\|\nabla F(\mathbf{x}_r)\right\|^2 \tag{70}$$

$$\tag{71}$$

Then we have

$$\mathbb{E}_r\|\mathbf{x}_{r+1}-\mathbf{x}_*\|^2\leq(1-\frac{\eta_cK\mu}{2})\|\mathbf{x}_r-\mathbf{x}_*\|^2+4\eta_c^2K^2\sigma_G^2+4\eta_c^2K^2\left\|\nabla F(\mathbf{x}_r)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2 \tag{72}$$

$$-\frac{2\eta_cK}{m}\sum_{i\in S_r}[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)]+\frac{2\eta_cL(1+K\eta_cL)}{m}\sum_{i\in S_r}\sum_{k\in[K]}\mathbb{E}_r\|\mathbf{x}_r-\mathbf{x}_{r,k}^i\|^2 \tag{73}$$

$$\leq (1-\frac{\eta_c K\mu}{2})\|\mathbf{x}_r-\mathbf{x}_*\|^2+4\eta_c^2K^2\sigma_G^2+4\eta_c^2K^2\left\|\nabla F(\mathbf{x}_r)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2 \tag{74}$$

$$-\frac{2\eta_c K}{m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]+2\eta_c KL(1+K\eta_c L)\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)+30K^2\eta_c^2\|\nabla F(\mathbf{x}_r)\|^2\right] \tag{75}$$

$$=(1-\frac{\eta_c K\mu}{2})\|\mathbf{x}_r-\mathbf{x}_*\|^2+4\eta_c^2K^2\sigma_G^2+4\eta_c^2K^2[1+15\eta_c KL(1+\eta_c KL)]\left\|\nabla F(\mathbf{x}_r)\right\|^2+\frac{K\eta_c^2}{m}\sigma^2 \tag{76}$$

$$-\frac{2\eta_c K}{m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]+2\eta_c KL(1+K\eta_c L)\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right]. \tag{77}$$

The last inequality is due to the upper bound of $\mathbb{E}_r\|\mathbf{x}_r-\mathbf{x}_{r,k}^i\|^2$. That is, for each client $i\in[M]$, we have $\mathbb{E}_r\|\mathbf{x}_r-\mathbf{x}_{r,k}^i\|^2\leq 5K\eta_c^2(\sigma^2+6K\sigma_G^2)+30K^2\eta_c^2\|\nabla F(\mathbf{x}_r)\|^2$.

For each step, we choose to use clients' update with probability q and server's update with probability $1-q$. Taking expectation and letting $\bar{\eta}=\frac{\eta_c K\mu}{2}=\frac{2\eta_s L\mu}{L+\mu}$, we have the following:

$$\mathbb{E}\|\mathbf{x}_{r+1}-\mathbf{x}_*\|^2\leq(1-\bar{\eta})\|\mathbf{x}_r-\mathbf{x}_*\|^2+4q\eta_c^2K^2\sigma_G^2+4q\eta_c^2K^2[1+15\eta_c KL(1+\eta_c KL)]\left\|\nabla F(\mathbf{x}_r)\right\|^2+\frac{qK\eta_c^2}{m}\sigma^2 \tag{78}$$

$$-\frac{2q\eta_c K}{m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]+2q\eta_c KL(1+K\eta_c L)\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right] \tag{79}$$

$$+(1-q)\eta_s\left(\eta_s-\frac{2}{L+\mu}\right)\left\|\nabla F(\mathbf{x}_r)\right\|^2+(1-q)\eta_s^2\sigma_s^2 \tag{80}$$

$$=(1-\bar{\eta})\|\mathbf{x}_r-\mathbf{x}_*\|^2+\bar{\eta}\left[\frac{4q\bar{\eta}}{\mu^2}\left(1+\frac{30L\bar{\eta}}{\mu}(1+\frac{2L\bar{\eta}}{\mu})\right)+(1-q)\frac{L+\mu}{2L\mu}\left(\frac{(L+\mu)\bar{\eta}}{2L\mu}-\frac{2}{L+\mu}\right)\right]\left\|\nabla F(\mathbf{x}_r)\right\|^2 \tag{81}$$

$$-\frac{4q\bar{\eta}}{\mu m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]+4q\eta_c^2K^2\sigma_G^2+\frac{qK\eta_c^2}{m}\sigma^2+2q\eta_c KL(1+K\eta_c L)\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right]+(1-q)\eta_s^2\sigma_s^2 \tag{82}$$

$$\leq(1-\bar{\eta})\|\mathbf{x}_r-\mathbf{x}_*\|^2+4q\eta_c^2K^2\sigma_G^2+\frac{qK\eta_c^2}{m}\sigma^2+2q\eta_c KL(1+K\eta_c L)\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right]+(1-q)\eta_s^2\sigma_s^2, \tag{83}$$

where $\left[\frac{4q\bar{\eta}}{\mu^2}\left(1+\frac{30L\bar{\eta}}{\mu}(1+\frac{2L\bar{\eta}}{\mu})\right)+(1-q)\frac{L+\mu}{2L\mu}\left(\frac{(L+\mu)\bar{\eta}}{2L\mu}-\frac{2}{L+\mu}\right)\right]\left\|\nabla F(\mathbf{x}_r)\right\|^2-\frac{4q}{\mu m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]\leq 0$. That is, $q\leq\frac{1}{1+\frac{\frac{4\bar{\eta}}{\mu^2}\left(1+\frac{30L\bar{\eta}}{\mu}(1+\frac{2L\bar{\eta}}{\mu})\right)G_3-\frac{4}{\mu}G_4}{\left(\frac{1}{L+\mu}-\frac{(L+\mu)^2\bar{\eta}}{4L^2\mu^2}\right)G_3}}$, where $G_3=\left\|\nabla F(\mathbf{x}_r)\right\|^2$ and $G_4=\frac{1}{m}\sum_{i\in S_r}\left[F_i(\mathbf{x}_r)-F_i(\mathbf{x}_*)\right]$.

Recursively applying the above and summing up the geometric series gives:

$$\mathbb{E}\|\mathbf{x}_R-\mathbf{x}_*\|^2\leq(1-\bar{\eta})^R\|\mathbf{x}_0-\mathbf{x}_*\|^2+\sum_{r=0}^{R-1}(1-\bar{\eta})^j\delta \tag{84}$$

$$\leq(1-\bar{\eta})^R\|\mathbf{x}_0-\mathbf{x}_*\|^2+\bar{\delta}, \tag{85}$$

where $\bar{\delta}=\frac{8q\eta_c K}{\mu}\sigma_G^2+\frac{2q\eta_c}{\mu m}\sigma^2+\frac{4qL(1+K\eta_c L)}{\mu}\left[5K\eta_c^2(\sigma^2+6K\sigma_G^2)\right]+(1-q)\frac{L+\mu}{2L\mu}\eta_s\sigma_s^2$.

Choosing $\bar{\eta}=\Omega(\frac{log(R)}{R})$, that is, $\eta_c=\Omega(\frac{log(R)}{R})$ and $\eta_s=\Omega(\frac{log(R)}{R})$, we have

$$\mathbb{E}\|\mathbf{x}_R-\mathbf{x}_*\|^2\leq\tilde{\mathcal{O}}(\frac{1}{R}).$$

□

Table 4. Test accuracy improvement (%) for SAFARI compared with FedAvg on CIFAR-10 ($s = 4$, $q = 0.8$). '-' means no statistical difference within 2% error bar.

DATASET SIZE	NON-I.I.D. INDEX (p)			
	10	5	2	1
500	-	-	3.45	7.14
5000	-	-	8.15	20.52
10000	-	2.93	9.32	23.86

Table 5. Test accuracy (%) for SAFARI ($p = 1$).

DATASET	s	CLIENT UPDATE PROBABILITY (q)			
		1.0 (FEDAVG)	0.8	0.6	0.4
MNIST	0	84.49	88.96	89.11	89.1
	2	71.58	88.11	88.06	88.27
	4	57.05	88.12	88.44	87.19
CIFAR-10	0	75.7	78.27	77.2	76.29
	2	64.56	77.13	75.17	75.08
	4	50.7	74.56	73.85	74.19

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 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 global learning learning rate 1.0, the local learning rate is 0.1, and the server learning rate for SAFARI is 0.1. 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 5000. 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 and the FedAvg algorithm. FedAvg is employed as the baselines to compare with our algorithm. To compare the effect of the collaboration from server, we add $[50, 100, 500, 1000]$ data to the server's side for MNIST and $[500, 5000, 10000]$ for CIFAR-10 and choose the client update probability $q = [0.4, 0.6, 0.8, 1.0]$. In the case of $q = 1.0$, our proposed algorithm SAFARI is equivalent to FedAvg.

B.2. Additional Experimental Results

In Table 4, 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 10000 data at server's side, the test accuracy increases by 23.86 %. 2) There is actually no improvement with these auxiliary data for nearly homogeneous case (e.g., $p = 10$), which is denoted by '-' in the table.

In Table 5, we show the test accuracy of SAFARI on MNIST and CIFAR-10 under different client update probability q. Note that when $q = 1.0$, SAFARI is equivalent to FedAvg. We can observe that even with a few server participation with a probability of 0.2, there is a non-negligible improvement in test accuracy.

B.3. Fashion-MNIST

We further run experiments with the Fashion-MNIST dataset, and the results are summarized as follows. These experiment results validate our theoretical findings.

Table 6. Test accuracy (%) for Fashion-MNIST dataset with different incomplete client participations.

s	0	30	60	90	120
Test accuracy	87.71 $\pm$ 0.09	86.17 $\pm$ 1.57	82.37 $\pm$ 5.8	80.71 $\pm$ 0.38	78.53 $\pm$ 3.05

Table 7. Test accuracy (%) for Fashion-MNIST dataset with different server's data. ($s = 90, M = 150, q = 0.8$.)

	FedAvg	SAFARI (1%)	SAFARI (10%)	SAFARI (20%)
Test accuracy	80 $\pm$ 0.38	82.0 $\pm$ 0.03	85.58 $\pm$ 1.05	85.14 $\pm$ 0.23

B.4. Ablation study about Random Initializations

We have run each experiment setting with five random initializations for the MNIST dataset and three random initializations for the CIFAR10 dataset. We report the mean and standard variance. The results are summarized in the following tables, which will also be added in the revision.

Table 8. Test accuracy (%) for MNIST dataset. (s is client sampling bias and p is non-i.i.d. index.)

		$p = 10$	$p = 5$	$p = 2$	$p = 1$
$s = 0$	FedAvg	92.67$\pm$ 0.05	89.70$\pm$ 0.23	86.04$\pm$ 0.61	84.60$\pm$ 0.99
	SAFARI	92.60$\pm$ 0.08	91.08$\pm$ 0.13	89.42$\pm$ 0.17	89.10$\pm$ 0.24
$s = 2$	FedAvg	92.64$\pm$ 0.07	89.13$\pm$ 0.28	86.34$\pm$ 0.92	71.66$\pm$ 0.74
	SAFARI	92.62$\pm$ 0.04	90.68$\pm$ 0.38	88.75$\pm$ 0.22	88.50$\pm$ 0.63
$s = 4$	FedAvg	92.58$\pm$ 0.03	88.76$\pm$ 0.13	77.82$\pm$ 0.22	57.09$\pm$ 0.04
	SAFARI	92.51$\pm$ 0.04	90.41$\pm$ 0.17	88.27$\pm$ 0.21	88.06$\pm$ 0.24

B.5. Ablation Study about Number of Clients

In addition to the important roles of p and s, we agree that m could serve as another influencing factor. Consequently, we have incorporated additional configurations for varying values of m, as delineated below. From these results, it is apparent that m has a negative effect on test accuracy in non-i.i.d. scenarios. As the non-i.i.d. degree increases, the negative influence of m becomes more pronounced.

Table 9. Test accuracy (%) for CIFAR10 dataset. (s is client sampling bias and p is non-i.i.d. index.)

		$p = 10$	$p = 5$	$p = 2$	$p = 1$
$s = 0$	FedAvg	81.48± 0.38	79.78± 0.34	78.36± 0.17	76.39± 0.60
	SAFARI	81.40± 0.45	80.13± 0.33	79.00± 0.25	78.34± 0.12
$s = 2$	FedAvg	80.11± 0.14	78.63± 0.33	77.14± 0.42	65.16± 0.56
	SAFARI	80.41± 0.10	79.13± 0.28	77.66± 0.10	77.21± 0.39
$s = 4$	FedAvg	78.16± 0.33	75.30± 0.22	67.78± 0.38	50.67± 0.25
	SAFARI	79.28± 0.21	78.27± 0.21	76.52± 0.16	75.18± 0.54

Table 10. Test accuracy (%) for MNIST dataset. (m is participated clients in training and p is non-i.i.d. index.)

		$p = 10$	$p = 5$	$p = 2$	$p = 1$
$m = 1$	FedAvg	92.11	65.63	58.65	49.86
	SAFARI	91.88	90.04	87.96	87.83
$m = 2$	FedAvg	92.40	83.28	67.03	53.75
	SAFARI	92.27	90.00	87.63	87.45
$m = 3$	FedAvg	92.51	87.19	75.14	55.76
	SAFARI	92.54	90.39	87.63	88.19
$m = 4$	FedAvg	92.55	88.60	78.52	56.42
	SAFARI	92.45	90.85	88.30	87.88
$m = 5$	FedAvg	92.62	88.81	77.81	57.05
	SAFARI	92.52	90.35	88.50	88.12
$m = 6$	FedAvg	92.58	88.20	77.42	57.12
	SAFARI	92.65	90.35	87.67	88.23

Table 11. Test accuracy (%) for CIFAR10 dataset. (m is participated clients in training and p is non-i.i.d. index.)

		$p = 10$	$p = 5$	$p = 2$	$p = 1$
$m = 1$	FedAvg	78.61	75.34	53.02	10.31
	SAFARI	79.80	78.70	75.90	63.04
$m = 2$	FedAvg	78.48	76.46	67.24	45.87
	SAFARI	79.15	77.90	76.75	72.93
$m = 4$	FedAvg	78.59	75.21	68.05	50.25
	SAFARI	79.98	77.56	76.60	75.51
$m = 5$	FedAvg	77.78	75.55	67.34	50.70
	SAFARI	79.47	78.48	76.66	74.56