Federated Learning Under Intermittent Client Availability and Time-Varying Communication Constraints

Mónica Ribero, Haris Vikalo, *Senior Member, IEEE*, and Gustavo de Veciana, *Fellow, IEEE*

***Abstract*—Federated learning systems facilitate the training of global models across large numbers of distributed edge-devices with potentially heterogeneous data. Such systems operate in resource constrained settings with intermittent client availability and/or time-varying communication constraints. As a result, the global models trained by federated learning systems may be biased towards clients with higher availability. We propose Federated Averaging Aided by an Adaptive Sampling Technique (F3AST), an unbiased algorithm that dynamically learns an availability-dependent client selection strategy which asymptotically minimizes the impact of client-sampling variance on the global model's convergence, enhancing performance of federated learning. The proposed algorithm is tested in a variety of settings for intermittently available clients operating under communication constraints, and its efficacy demonstrated on synthetic data and realistically federated benchmarking experiments using CIFAR100 and Shakespeare datasets. We report up to 186% and 8% accuracy improvements over FEDAVG, and 8% and 7% over FEDADAMon CIFAR100 and Shakespeare, respectively.**

***Index Terms*—Edge learning, distributed learning, federated learning, resource management, communication efficiency.**

I. Introduction

FEDERATED learning (FL) has emerged as an attractive framework in edge learning to train models when the data is distributed among edge devices and must remain local due to resource constraints and/or privacy concerns. The edge-device networks in FL could comprise millions of clients [1] whose feedback might include model updates that are on the order of 100 Mb. For example, neural network for image recognition tasks VGG-16 [2] has 160 M parameters and weights resulting in updates of size 526Mb when using 32 b encoding.

In the original Federated Averaging algorithm (FEDAVG) [3], as well as more recent approaches including SCAFFOLD [4], Federated Adaptive Optimization [5], FEDDYN [6] and FedProx [7], a server selects a random subset of clients and which will participate in updating a global model by training on local data. The server aggregates the clients' updates to produce a new global model, broadcasts it to the clients, and a new round of training begins; this procedure is repeated until convergence.

The potentially large amount of communication between the clients and the server makes sub-selection policies that reduce data traffic imperative. Additionally, one might want to judiciously account for differences in clients' availability patterns. Such patterns reflect inherent biases that may adversely affect learning goals, e.g., some devices may be more willing to participate as they may be less energy or bandwidth constrained. Indeed, one of the biggest gaps between theory and practice of FL is due to biases in sampling of edge-devices resulting from heterogeneous, possibly stochastic, on-and-off availability and communication constraints [1], [8], [9]. For example, in cross-device settings including mobile device systems [3], a vast number of client devices [10] with limited communication and power resources [1] *intermittently* connects to a central server to help optimize a global objective. Existing FL algorithms typically ignore intermittency and assume that the participating client devices are always available and thus can be tasked with performing a model update at any time [3], [6], [11], [12]. If not addressed by the system design, time-varying communication constraints and intermittent client availability (due to battery and other device-specific limitations) may cause significant degradation of the learned model performance [8], [11], [13].

To illustrate the potential severity of the problem described above, and preview the contributions of this paper, consider the following simple example. Let c_1 and c_2 be two clients with distinct data distributions. A server aims to optimize the function $F(\mathbf{w}) = p_1 F_1(\mathbf{w}) + p_2 F_2(\mathbf{w})$ over $\mathbb{R}^p$, where F_1 and F_2 denote the loss functions at clients c_1 and c_2, respectively, and for simplicity $p_1 = p_2 = 1/2$. We shall consider a model for the clients' intermittent availability characterized by the joint distributions given in Table I, where A_i is a binary random variable indicating whether client c_i is available. In this model, clients' availabilities in a given round are independent, with $P(A_1 = 1) = 0.375$ while $P(A_2 = 1) = 0.8$. Note the client availabilities are assumed to be independent across rounds. Suppose there is a communication constraint which restricts the server to sampling at most a single client each round. The server must thus choose a possibly client state dependent policy

Manuscript received 31 May 2022; revised 7 October 2022; accepted 7 November 2022. Date of publication 24 November 2022; date of current version 17 February 2023. This work was supported in part by the National Science Foundation under Grant 2148224 and in part by OUSD R&E, NIST, and Industry Partners as specified in the Resilient & Intelligent NextG Systems (RINGS) Program. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Xu. *(Corresponding author: Mónica Ribero.)*

The authors are with Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78703 USA (e-mail: mribero@google.com; hvikalo@ece.utexas.edu; deveciana@utexas.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTSP.2022.3224590, provided by the authors.

Digital Object Identifier 10.1109/JSTSP.2022.3224590

TABLE I
CLIENT AVAILABILITY MODEL: THE AVAILABILITY IS INDEPENDENT ACROSS TIME AND CLIENTS

	$A_2 = 1$	$A_2 = 0$	Marginal
$A_1 = 1$	0.3	0.075	0.375
$A_1 = 0$	0.5	0.125	0.625
Marginal	0.8	0.2	

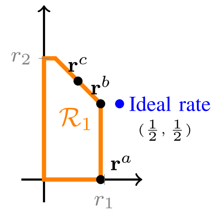

Fig. 1. The region of achievable long-term participation rates under the client availability model in Table I.

for selecting the clients in each round. Each such policy would achieve certain long-term client participation rates across the rounds, denoted $\mathbf{r} = (r_1, r_2)$. For example, under the model in Table I, the set of achievable long-term participation rates across all possible policies is given by the region $\mathcal{R}_1$ shown in Fig. 1. Given the communication constraints, it is not possible to achieve the full client participation rates of $\mathbf{r} = (1, 1)$ because clients are not always available and we can only sample one client in each round. However, $\mathbf{r}^a = (0.375, 0)$ is achievable by using the state-dependent deterministic policy which selects c_1 whenever c_1 is available and never selects c_2. Alternatively, $\mathbf{r}^b = (0.375, 0.5)$ is also achievable by selecting c_1 whenever c_1 is available, and c_2 when only c_2 is available. A naive selection policy that samples from available clients with probability proportional to $p_i = \frac{1}{2}$ in hope of achieving "ideal" participation rate $(\frac{1}{2}, \frac{1}{2})$ [12] would actually result in client c_1 participating at a rate of

$$r_1^c = P(A_1 = 1, A_2 = 0) + \frac{P(A_1 = A_2 = 1)}{2} = 0.225.$$

Analogously, the long-term participation rate of client c_2 under the same naive selection policy is $r_2^c = 0.65$.

As demonstrated in Section III (Theorem III.5), improper choice of the long-term participation rate $\mathbf{r}$ injects bias and variance into the global model. Therefore, selecting an "appropriate" rate is of fundamental importance; yet, as illustrated above, intermittent client availability and communication constraints present several previously overlooked challenges: (i) determining the long-term participation rate $\mathbf{r}^* \in \mathcal{R}$ which is best in terms of its impact on the convergence of federated learning, and (ii) design of a client selection policy that achieves rate $\mathbf{r}^*$. These are particularly demanding because the (possibly correlated) clients' availability patterns are unknown and, therefore, $\mathcal{R}$ is unknown.

The main contribution of this paper is learning to sample clients in large edge-device FL networks with heterogeneously distributed data and intermittent client availability. In particular we introduce F3AST, a federated learning algorithm that also learns how to adapt its client sampling strategy to *unknown* client availability statistics and adapts to time-varying communication constraints. F3ASTis shown to be asymptotically optimal (see Theorem III.3) as its long-term participation rate converges to the value minimizing a bound on the global model variance over the space of achievable rates. Remarkably, F3ASTaccomplishes this *with no prior knowledge of the communication constraints or clients' availability models*. To our knowledge, this is the first work to *formally* address client intermittency and system capacity variability in federated learning with data-heterogeneity, and the first work to propose a method to *learn how to select clients* while pursuing a shared (global) model within the federated learning framework.

Extensive experimentation on realistic tasks and data. F3ASTis tested on three benchmark datasets: (i) Synthetic(1,1) [14], a widely used heterogeneous synthetic dataset for softmax regression [7]; (ii) a realistically federated version of CIFAR100 [5]; and (iii) Shakespeare [3]. We demonstrate that in learning highly non-linear models F3ASTexhibits more stable convergence and considerably higher accuracy than state-of-art algorithms. Moreover, F3AST's selection and aggregation method is readily combined with the existing optimization techniques designed to address system's constraints, allowing those methods to take advantage of F3AST's policies: experiments confirm that incorporating F3ASTreduces bias of algorithms that do not compensate for client selection uncertainties, and demonstrate much more stable descent trajectories to the optimum even in highly time-varying environments.

II. BACKGROUND AND RELATED WORK

A. Federated Learning

Given a set $\mathcal{U}$ with N clients, each having n_k data samples, a federated learning system is concerned with solving

$$\min_{\mathbf{w} \in \mathbb{R}^p} \mathbb{E}_{k \sim \mathcal{P}}[F_k(\mathbf{w})], \tag{1}$$

where $F_k(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_k}[f_k(\mathbf{w}; \xi)]$ denotes the loss function of client k and $\mathcal{P}$ is the distribution over users. The generalized FEDAVG algorithm, FEDOPT [5], interactively learns the global model by *randomly selecting* at time t a subset of clients S_t to locally optimize their objective function[1] starting from the initial model $\overline{\mathbf{w}}^t$, and communicate their updates $\mathbf{v}_k^{t+1}$ to the server. Then, the server *aggregates* the received updates $(\mathbf{v}_k^{t+1})_{k \in S_t}$ to produce the global update Δ^{t+1} and generate a new global model $\overline{\mathbf{w}}^{t+1}$. For the remainder of the paper we refer to the process of going from $\overline{\mathbf{w}}^t$ to $\overline{\mathbf{w}}^{t+1}$ as a (single) round of FL. The process is repeated with the aim of finding accurate global model. Heterogeneity due to generally non-i.i.d. and unbalanced data available to different clients emerges as one of the main challenges in federated learning, and thus the choice of a client sampling scheme heavily impacts convergence of the global model. Several authors have addressed this problem, proposing different optimizers [5] and client sampling strategies [11], [12], but all showed convergence (or near convergence) only under the strong assumption of being able to sample any client at any time. A technique that deals with heterogeneity by allowing each client to learn a personalized model was proposed in [15].

[1] E.g., the original FEDAVG algorithm coordinates E epochs over training data.

Several approaches to addressing communication constraints in FL systems have recently been proposed in literature, including strategies aiming to reduce client communication rate via model compression [16], [17], [18], [19], [20], [21]. These are orthogonal to our work since we focus on settings where the clients are intermittently available.

a) Client availability: We assume that, at any time, the set of available clients is random and non-empty, and that the constraint on the number of clients selected to train and provide model updates to the server generally varies over time. Note that, in such scenarios, applying recent stateful optimization techniques such as SCAFFOLD [4] and FEDDYN [6] is challenging due to hardware constraints and high number of participating devices [9]. Prior work has investigated client availability under restrictive conditions such as block-cyclic data characteristics [8], [11], assumed i.i.d. availability across clients that act as stragglers [7], or produced biased models [22]. However, availability is much more difficult to model in practice. Although certain patterns such as day/night are cyclic, there also exist various other non-cyclic client or cluster-specific patterns that affect some clients more than others, including access to a power source and the available communication bandwidth.

B. Client Sampling and Averaging

Since the number of clients in federated learning systems can be extremely large [23], [24], only a relatively small subset of them is tasked with training in each round. Data heterogeneity and communication constraints have inspired several strategies for selecting clients from the available pool. Some of those techniques take into account the proportion of data at each client, which we denote by p_k [12]. Alternative strategies apply active learning ideas to client selection and select those that are more promising according to some metric, e.g., choose clients with the largest magnitude of the updates [13], [25], [26], [27], [28] or those with the highest loss [11].

Another line of related prior work has been focused on investigating model aggregation strategies. In [13], the authors assume that stochastic optimization updates approximately follow a stationary stochastic process, and cast the model aggregation as an estimation problem. An alternative aggregation strategy is to form an unweighted average of the updates [11]; however, this leads to a biased model and large variance. Other approaches trade communication and memory for stable convergence [4], or replace missing updates with the previous model [26], [27]. In the centralized setting, importance sampling has been used to optimally aggregate SGD updates [29], [28].

Previous works on client sampling in FL systems do not provide formal convergence guarantees for settings where the clients are intermittently available. Work in [8], [11] study effects of cyclically alternating client availability and propose a sampling strategy empirically shown to improve over random sampling; however, the resulting models may be biased and their performance under non-cyclic availability patterns is unclear.

Alternatively, asynchronous methods address client selection under system heterogeneity with a fixed selection policy determined by clients training speed: updates are incorporated individually as they arrive at the server [30] or, when there are privacy concerns, they are aggregated in buffers [31].

III. METHODS

In this section we present and analyze a novel framework for selecting and aggregating intermittently available clients in federated learning systems that operate under time-varying communication constraints. In such settings, the contribution each client makes to the federated averaging process depends on how often the client is selected to provide an update – i.e., on the *long-term client participation rate*. We start by characterizing the set $\mathcal{R}$ of achievable long-term client participation rates subject to communication and client availability constraints. Then, we introduce F3AST,an algorithm that dynamically learns clients' long-term participation rate and improves the convergence of federated learning by reducing the model bias and minimizing variance introduced by sampling intermittently available clients. The omitted proofs can be found in the appendix.

A. Preliminaries

Communication constraints and intermittent client availability: Consider a FL system in which a random subset of clients $\mathbf{A}_t$ is available/responsive at time t; here $(\mathbf{A}_1, \mathbf{A}_2, \dots) = (\mathbf{A}_t)_t$ form a discrete-time stochastic process with a finite state space $\mathcal{A} = 2^{\mathcal{U}}$, i.e., the collection of all possible subsets of the set of users $\mathcal{U}$. Communication constraints restrict the possible subsets of clients that can be chosen to participate during a training round; we let $\mathbf{C}_t$ denote the (random) collection of the available clients sets that meet communication constraints at time t and denote its state space by $\mathcal{C}$; given $\mathbf{A}_t = A$, a realization $\mathbf{C}_t = C$ corresponds to a collection of subsets of A, i.e., $C \subseteq 2^A$. For convenience, in the remainder of the paper we refer to $\mathbf{C}_t$ as the system *configuration*.

To illustrate the use of the introduced notation, consider the communication-constrained setting where the number of clients allowed to participate in training round t is no more than (possibly random) K_t. Given a realization of the set of clients available at time t, $\mathbf{A}_t = A$, and the aforementioned communications constraint $K_t = k \in \mathbb{N}$, the collection of feasible client sample sets S that the server may choose to include in the training round is

$$C = \{S \subset A : |S| \leq k\}.$$

If the communication constraints are not time-varying, i.e., if $K_t = k$ almost surely for all t, we are back in the traditional FEDAVG cross-device setting.

Assumption 1: The sequence of random collections of feasible client sampling sets $(\mathbf{C}_t)_t$ forms a discrete-time irreducible Markov chain with a finite state space $\mathcal{C} \subseteq 2^{\mathcal{U}}$ and a stationary distribution $\pi = (\pi(C), C \in \mathcal{C})$.

Assumption 1 significantly relaxes assumptions typically made when analyzing convergence of state-of-the-art FL algorithms, e.g., the much stronger assumptions of all users having unlimited availability [3], [4], [6], [7], [32], [33] or a deterministic block-cyclic availability [8]. Assumption 1 captures various realistic settings including that of home devices available with a given probability - not necessarily uniform across clients - throughout the day. Our experimental results showcase that in several realistic settings which meet Assumption 1, our method provides significant performance improvements while other techniques fail to adapt to unknown availability models.

b) Static configuration-dependent client sampling policies: Communication constraints and intermittent client availability restrict the space of admissible long-term client participation rates. Indeed, it is unrealistic to expect being able to sample an arbitrary collection of k clients at time t with pre-specified probabilities $\mathcal{P} = (p_i : i = 1 \ldots N)$ since some of those k clients may be unavailable, and/or time horizon is not long enough to achieve certain rate (see examples in Section I). To characterize *achievable* long-term participation rates we introduce the following concepts.

Recall that for a given configuration of communication constraints and client availabilities there exists an associated collection of feasible client sample sets C that a sampling policy can choose from. We define a *static configuration-dependent client sampling policy* as follows.

Definition III.1: For each C, let $f_{C,S} \geq 0$ denote the probability of selecting the subset of clients $S \in C$, where $\sum_{S \in C} f_{C,S} = 1$. If we denote $\boldsymbol{f}_C = (f_{C,S}, S \in C)$, then $\boldsymbol{f} := (\boldsymbol{f}_C, C \in \mathcal{C})$ specifies a static configuration-dependent sampling policy selecting clients over different communication/availability configurations.

Let $\mathcal{F}$ denote the set of possible static configuration-dependent client sampling policies. Under the above model, the long-term client participation rate can be expressed as

$$\mathbf{r}^{\boldsymbol{f}} = \sum_{C \in \mathcal{C}} \pi(C) \sum_{S \in C} f_{C,S} \mathbb{1}_S, \tag{2}$$

where $\mathbb{1}_S$ is an N-dimensional binary indicator vector whose i^{th} entry is 1 if the i^{th} client is in S, and is 0 otherwise. One can interpret the i^{th} component of vector $\mathbf{r}^{\boldsymbol{f}}$ as the fraction of time the i^{th} client is selected by the server.

Finally, we define the long-term client participation rate region as the set of all possible long-term participation rate vectors $\mathbf{r}^{\boldsymbol{f}}$, i.e., $\mathcal{R} := \{\mathbf{r}^{\boldsymbol{f}} | \boldsymbol{f} \in \mathcal{F}\}$.

Lemma III.2: The long-term client participation rate region $\mathcal{R} = \{\mathbf{r}^{\boldsymbol{f}} | \boldsymbol{f} \in \mathcal{F}\}$ is a subset of the simplex in the N-dimensional Euclidean space, and a closed convex set.

Proof: The lemma follows from the fact that $\mathcal{R}$ is a linear image of all possible $\boldsymbol{f}$, a closed bounded convex set. $\mathcal{R}$ defines the region of achievable participation rates. We rely on this lemma to prove convergence of our algorithm to the optimal rate in Theorem III.3. ■

B. F3AST: Minimizing the Sampling Variance

Here we formally introduce an algorithm that learns a client selection policy which ensures that the resulting long-term client participation rate converges to a value minimizing

$$H(\mathbf{r}) := \begin{cases} \sum_{k=1}^{N} \frac{p_k}{r_k} & \text{client availability is} \\ & \text{positively correlated,} \\ \sum_{k=1}^{N} \frac{p_k^2}{r_k} & \text{otherwise,} \end{cases} \tag{3}$$

where positive correlation between availability of clients i and j implies that an event of client i being available increases the probability of client j being available. Whether client availability is positively correlated or not depends on application domain; if the nature of availability correlation is unknown, minimizing $\sum \frac{p_k}{r_k}$ remains a meaningful option since this objective provides a bound on the variance in all cases. It is readily shown that $H(\mathbf{r})$ bounds the variance induced in the global model by the selection policy with rate $\mathbf{r}$. For the ease of exposition we postpone that discussion to Section III-B2 in favor of first presenting our proposed algorithm.

F3AST(**F**ederated **A**veraging **A**ided by an **A**daptive **S**ampling **T**echnique), is presented as Algorithm 1. Formally, F3ASTaims to find a configuration-dependent client sampling strategy $\boldsymbol{f}_{\text{F3AST}}$ such that its long-term client participation rate $\mathbf{r}_{\text{F3AST}}$ approximates the optimal achievable strategy $\mathbf{r}^* \in \text{argmin}_{\mathbf{r} \in \mathcal{R}} H(\mathbf{r})$. To accomplish this, F3ASTfirst initializes $\mathbf{r}(0)$ arbitrarily (line 1). At each round t, with $\mathbf{C}_t = C_t$, selecting set $S \in C_t$ implies a contribution to the participation rate of 1 for every client $k \in S$. Whether or not selecting set S brings $r(t)$ closer to $\mathbf{r}^*$ can be computed by estimating the marginal utility of S using the gradient of $H(\mathbf{r})$. Thus, we select S_t (line 5) as

$$S_t \in \arg\max_{S \in C_t} -\nabla H(\mathbf{r}(t)) \cdot \mathbb{1}_S. \tag{4}$$

In general, (4) is a combinatorial optimization problem; in the federated learning systems with a time-varying bound on the number of clients K_t that can be selected ($K_t \geq 0$), (4) reduces to the discrete optimization problem of greedily selecting K_t available clients with the largest entries of $-\nabla H(\mathbf{r}(t))$. Optimality of the greedy approach follows because the objective is an additive set function.

Next, the rate is updated to reflect the selection made in the latest iteration of the sampling scheme. This is done by forming an exponentially smoothed average of the past sampling rates (line 6),

$$\mathbf{r}(t+1) = (1-\beta)\mathbf{r}(t) + \beta \mathbb{1}_S, \tag{5}$$

where $\beta > 0$ is a fixed small parameter, set to $\beta = O(1/T)$ for convergence purposes. After having selected clients S_t, the server broadcasts the current model and the selected clients perform local updates using a local optimization procedure CLIENTOPT($\overline{\mathbf{w}}^t$). Finally, the server uses $\mathbf{r}(t)$ to produce an unbiased global model $\overline{\mathbf{w}}^{t+1}$ with estimator Δ^{t+1} (lines 9–10).

a) Beyond FEDAVG: F3ASTmodifies two crucial steps in FEDAVG: client sampling and model updates aggregation. This makes it suitable to work in combination with other FL algorithms like SCAFFOLD [4], AFL [33], FEDPROX [7], FEDDYN [6], and more generally FEDOPT [5]. These methods' theoretical guarantees are provided under the "all clients are available" assumption, implying that they assume an unrealistic fixed sampling policy which introduces bias to the model. Our proof

Algorithm 1: F3AST: Federated Averaging Aided by an Adaptive Sampling Technique

```
Input: Server parameters: learning rate schedule ${\eta_t}_{t=1}^T$,
    the number of global rounds $T$, the number of clients
    per round $K_t$, the number of client local updates $E$,
    $\beta = O(1/T)$
Output: Global model $\overline{\mathbf{w}}_T$
initialize $\overline{\mathbf{w}}_0 \in \mathbb{R}^p$ arbitrarily, initialize $\mathbf{r}(0)$ arbitrarily
for $t = 1 \to T$ do
    $C_t \leftarrow$ feasible client sets at time $t$
    $S_t \in \text{argmin}_{S \in C_t} \nabla H(\mathbf{r}_t) \mathbb{1}_S$
    $\mathbf{r}(t) = (1-\beta)\mathbf{r}(t-1) + \beta \mathbb{1}_S$
    for Clients $k \in S_t$, in parallel do
        $\mathbf{v}_k^{t+1} \leftarrow \text{CLIENTOPT}(\overline{\mathbf{w}}^t, E \text{ steps}, \eta_t)$
    end for
    $\Delta^{t+1} \leftarrow \sum_{k \in S_t} \frac{p_k}{r_k(t)} \mathbf{v}_k^{t+1}$
    $\overline{\mathbf{w}}^{t+1} \leftarrow \text{SERVEROPT}(\overline{\mathbf{w}}^t, \Delta^{t+1})$
end for
```

extends to those settings by modifying accordingly sampling and aggregation to the asymptotically learned $\mathbf{r}$ as long as the ℓ_2-norms of clients' model updates are uniformly bounded – an assumption already made by the above methods.

1) Asymptotic Optimality

Below we show that as $\beta \downarrow 0$, the selection policy rate converges to the value that optimizes $H(\mathbf{r})$ and, therefore, reduces the model variance. To this end, consider the discrete time Markov process $S^\beta(t) = (\mathbf{r}^\beta(t), \mathbf{C}_t)$ indexed by the value of β defined in Eq. (5), with $\beta \downarrow 0$ along sequence $\mathcal{B} = \{\beta_j\}_{j \in \mathbb{N}}$. $S^\beta(0)$ and the probability law of $\mathbf{C}_t$ describing the availability model are fixed for all $\beta \in \mathcal{B}$. The speed of convergence is discussed in the appendix.

Theorem III.3: Let $\mathbf{r}^\beta(t)$ be the rate determined by Algorithm 1. Let $V \subset \mathbb{R}_+^N$ be a bounded set, $\epsilon > 0$, and let $\mathbf{r}^*$ denote the minimizer of the variance function $H(\mathbf{r})$ over $\mathcal{R}$. Then for $T > 0$, depending on ϵ and V,

$$\lim_{\beta \downarrow 0} \sup_{\mathbf{r}^\beta(0) \in V, t > T/\beta} P[\|\mathbf{r}^\beta(t) - \mathbf{r}^*\| > \epsilon] = 0.$$

2) Bounding the Global Model Variance

We start by analyzing a fixed arbitrary stochastic policy $\boldsymbol{f}^{\mathbf{r}}$ achieving rate $\mathbf{r}$ and use the result to demonstrate that $H(\mathbf{r})$ reflects the model variance induced by $\boldsymbol{f}^{\mathbf{r}}$. Let us introduce several assumptions regarding clients' loss functions $F_k(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_k}[f_k(\mathbf{w}; \xi)], 1 \leq k \leq N$; these assumptions are commonly encountered in the federated learning literature [11], [12].

Assumption 2. [Smoothness and strong convexity]: $F_1, \ldots, F_N$ are $L-$smooth and $\mu-$strongly convex functions, meaning that for all v and w, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + \nabla F_k(\mathbf{w})^T(\mathbf{v} - \mathbf{w}) + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$ and $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + \nabla F_k(\mathbf{w})^T(\mathbf{v} - \mathbf{w}) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$, respectively.

Assumption 3: [Bounded variance]: Let ξ be a data point that client k samples from distribution $\mathcal{D}_k$. Then $\mathbb{E}_{\xi \sim \mathcal{D}_k}[\|\nabla f_k(\mathbf{w}, \xi) - \nabla F_k(\mathbf{w})\|_2^2] \leq \alpha_k^2, k = 1, \ldots, N$.

Assumption 4: [Bounded stochastic gradients] The expected norm of the stochastic gradients of f_k is uniformly bounded, i.e., $\mathbb{E}_{\xi \sim \mathcal{D}_k}[\|\nabla f_k(\mathbf{w}, \xi)\|_2^2] \leq G^2, k = 1, \ldots, N$.

Assumption 2 holds in a number of scenarios of interest, including ℓ_2-regularized linear and logistic regression, and classification with softmax function. Assumptions 3 and 4 are common in state-of-the-art distributed learning literature [34], [35], [36], [37]. Note that while we rely on the above assumptions when analyzing the performance, experimental results demonstrate that our sampling techniques work very well in more general settings involving highly non-linear models such as convolutional and recurrent neural networks trained on realistic datasets.

The following lemma introduces and analyzes $\sigma_t^2(\boldsymbol{f}^{\mathbf{r}})$, the client sampling variance under sampling policy $\boldsymbol{f}^{\mathbf{r}}$.

Lemma III.4: Suppose Assumptions 1-4 hold. Let $\mathbf{r} \in \mathcal{R}$ be an achievable long-term client participation rate under the system configuration determined by distribution π, and $\boldsymbol{f}^{\mathbf{r}}$ denote a static configuration-dependent selection policy achieving $\mathbf{r}$. Define the client sampling variance $\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) := \frac{1}{\eta_t^2}\mathbb{E}_{S_t}[\|\Delta^t - \overline{\mathbf{v}}^t\|^2]$ where $\overline{\mathbf{v}}^t = \sum_{k=1}^N p_k \mathbf{v}_t^k$ is the update at time t with full client participation. Then

$$\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) \leq 4E^2G^2 \left(\sum_{k=1}^N \frac{p_k}{r_k} - 1 \right). \quad (6)$$

Furthermore, if client availabilities are uncorrelated or negatively correlated, then there exists a policy $\boldsymbol{f}^r$ such that

$$\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) \leq 4E^2G^2 \left(\sum_{k=1}^N \frac{p_k^2}{r_k} + \sum_{k=1}^N p_k^2 \right). \quad (7)$$

Theorem III.5: Instate the settings of Lemma III.4. Let $\mathbf{w}^*$ denote the solution to the optimization problem (1), and $L, \mu = O(1)$. Define $\gamma = \max\{8\frac{L}{\mu}, E\}$, and assume learning rate $\eta_t = \frac{2}{\mu(\gamma + t)}$. Then by setting CLIENTOPT to SGD and SERVEROPT$(\overline{\mathbf{w}}^t, \Delta^{t+1}) = \overline{\mathbf{w}}^t + \Delta^{t+1}$, the model $\overline{\mathbf{w}}^T$ produced with policy $\boldsymbol{f}^{\mathbf{r}}$ after T steps satisfies

$$\mathbb{E}[F(\overline{\mathbf{w}}^T)] - F^* = O\Bigg(\frac{1}{TE + \gamma} \big(\|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \sum_{k=1}^N p_k^2 \alpha_k^2 + 6\Gamma + 8(E-1)^2G^2 + \sigma_T^2(\boldsymbol{f}^{\mathbf{r}}) \big) \Bigg), \quad (8)$$

where $\Gamma = F^* - \sum_{k=1}^N p_k F_k^*$ denotes the local-global objective gap[2], F^* and F_k^* are the minimum values of F and F_k, respectively, and $\sigma_T(\mathbf{r})$ (Lemma III.4) captures the variance induced by client sampling.

[2] Local-global objective gap quantifies data heterogeneity: for i.i.d. data, $\Gamma \to 0$ as the number of samples grows, while a large Γ indicates a high degree of heterogeneity [11], [12].

Remark III.6: Proof of Theorem III.5 follows the line of argument similar to that in the analysis of FL algorithms convergence [5], [12], [35]. The first term in the parenthesis in eq. (8) captures the effect of initialization, while the second term reflects the variance of stochastic gradients. The remaining terms are tied to the inherent challenges of data heterogeneity in FL.

From static to dynamic policies: Aiming to circumvent the requirement for having access to (generally unavailable) system configuration information that applies to static selection policies, we proceed by combining expressions (6) and (7) (ignoring constant terms) to define

$$H(\mathbf{r}) := \begin{cases} \sum_{k=1}^{N} \frac{p_k}{r_k} & \text{client availability is positively correlated,} \\ \sum_{k=1}^{N} \frac{p_k^2}{r_k} & \text{otherwise.} \end{cases}$$

Minimizing $H(\mathbf{r})$ over $\mathbf{r}$ reduces the upper bound on the model variance stated in Lemma III.4. It follows from (2) that $r_k \leq 1$, and thus it can readily be shown that $r_k = 1$ for all k minimizes $H(\mathbf{r})$ over $[0,1]^N$. However, it is possible that for a given configuration of the system this long-term client participation rate is not achievable, i.e., $\mathbf{1} \notin \mathcal{R}$ (the set of feasible $\mathbf{r}$'s defined in Lemma III.2). Recalling the example in Section I, minimization of $H(\mathbf{r})$ over $\mathcal{R}$ is also difficult because it requires knowledge of the achievable long-term client participation rate region $\mathcal{R}$, determined by the distribution π defined in Assumption 1. Finally, even if $\mathcal{R}$ is known, the resulting policy $\boldsymbol{f}^{\mathbf{r}}$ will likely be client and set dependent, thus rendering the problem challenging due to an exponential number of variables and unknown parameters. These are precisely the obstacles that F3ASTovercomes by learning a selection policy which is asymptotically optimal in terms of minimizing $H(\mathbf{r})$, and guaranteeing convergence to a long-term participation rate minimizing $H(\mathbf{r})$ over $\mathcal{R}$.

Rapid mixing time: Convergence rate of $\mathbf{r}(t) \rightarrow \mathbf{r}^*$ depends on the properties of the Markov chain specified by the availability process. Concretely, we have the following known theorem (see, e.g., [38] for a proof and details).

Theorem III.7. [Convergence Theorem]: Let P be the transition matrix of a system configuration satisfying Assumption 1 with stationary distribution π. Then there exists a constant $\alpha \in (0,1)$ and $C > 0$ such that

$$\max_{C \in \mathcal{C}} \|P^t(C,\cdot) - \pi\|_{TV} \leq C\alpha^t.$$

The above result shows that in practice one needs $t_0 \geq \frac{\log \epsilon}{\log \alpha}$ iterations to achieve a stationary rate $\mathbf{r}$ up to an error ϵ to the stationary distribution. Further, even if the rate is not constant during the early iterations, this result demonstrates that after a burn-in the rate will stabilize and the asymptotic convergence rate of $O(1/TE)$ will not be affected.

IV. Experiments

a) Datasets and models: We test our model on three well-known federated datasets. First, a synthetic heterogenous dataset Synthetic(1,1) for softmax regression, introduced in [14] and widely used in the FL community [7], [11], [12]. Second, a recurrent neural network with 1 M parameters for the next character prediction task on the Shakespeare dataset [3], a language modelling dataset with 725 clients, each one a different speaking role in each play from the collective works of William Shakespeare. Third, CIFAR100 with the partition introduced in [5], utilizing Latent Dirichlet Allocation in order to generate a realistic heterogenous distribution. We train ResNet-18, replacing batch with group normalization, a modification that has shown improvements in federated settings [39]. Our code is available on Github[3]. Number of clients and total number of samples is summarized in Table II.

TABLE II
DATASETS

Dataset	Users	Samples
Synthetic	100	60 K
CIFAR100	500	50 K
Shakespeare	715	16 K

b) Availability models: We perform tests on five realistic availability models described below. To our knowledge, there exist no public databases with real availability patterns; Smartphone's model [1] is inspired by realistic data. All models are motivated by practical federated learning systems:

1) *Always:* Baseline model, clients are always available.
2) *Scarce:* Independent and homogeneous availability across clients and time with probability $q = 0.2$.
3) *Home-devices:* Independent availability across clients and time with probability $q_k = T_k/B$, where $T_k \sim \text{lognormal}$ and $B = \max_k T_k$.
4) *Smartphones:* Sine-modulated *Home-devices* model, $q_{k,t} = f_t q_k$, where q_k is defined in the *Home-devices* model and f_t denotes a sinusoidal time-dependent availability (see [1]).
5) *Uneven:* Each client's availability is inversely proportional to its dataset size, $q_k \propto 1/p_k$.

We split each client's dataset into training and validation sets. We assume the distribution $\mathcal{P}$ over users is determined by the fraction of data they possess. In the following, we first fix the communication constraint to select $K = 10$ clients in each round and compare different methods across availability models. We then proceed by exploring the effect of varying K. We include further details on the experimental setup in Appendix C-A. We implement our models using the Tensorflow-Federated API [40].

c) Baselines: First, we compare our algorithm with two availability-agnostic methods: (i) FedAvg, a standard baseline, and (ii) FedAdam, which achieves state-of-the-art performance in the considered benchmark tasks [5]. Both methods sample available clients with normalized probabilities p_k, but FedAvg uses SGD as the server optimizer while FedAdamuses Adam. For a fair comparison, we implement both methods and compare with their availability-aware versions wherein we incorporate our proposed sampling and aggregation step.

Second, we test against a state-of-the-art algorithm, Power-of-Choice (PoC) [11], a method that, although agnostic to the

[3] [Online]. Available: https://github.com/mriberodiaz/f3ast

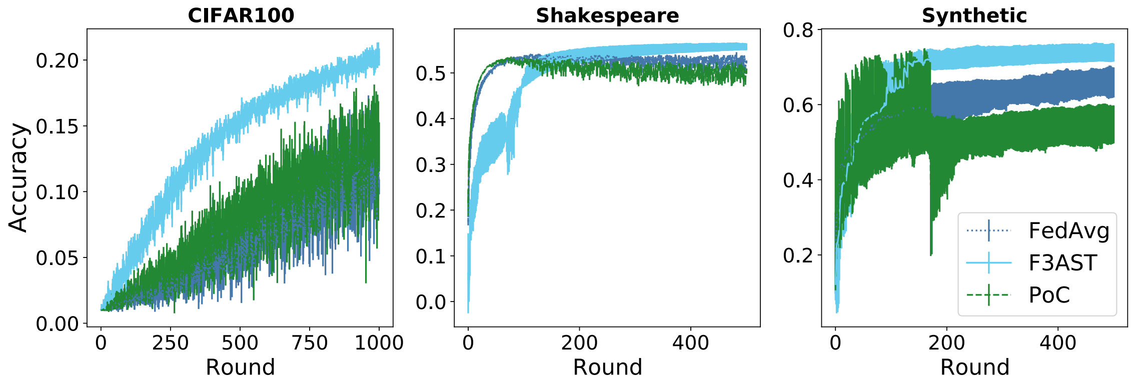

Fig. 2. Test per-sample accuracy (averaged over three runs) for different client sampling and aggregation schemes under *HomeDevice* availability model. We observe that F3ASTconsistently outperforms FedAvg and PoC. Further, F3ASTstabilizes while FedAvg and PoCare unable to adapt to the time-varying environment.

TABLE III
Test Sample Accuracy of All Methods, and Relative Improvements of F3ASTOver FedAvg and F3AST+ Adam Over FedAdam, for Different Availability Models (columns) on CIFAR100 (1000 Rounds) and Shakespeare (500 Rounds)

		Availability models				
		Always	Scarce	HomeDevice	Uneven	SmartPhones
CIFAR100	FedAvg	0.141	0.096	0.111	0.072	0.142
	F3AST	0.198 (+40%)	0.201 (+109%)	0.208 (+87%)	0.206 (+186%)	0.201 (+42%)
	FedAdam	0.262	0.288	0.302	**0.282**	0.298
	F3AST + Adam	**0.271** (+3%)	**0.308** (+7%)	**0.324** (+7%)	**0.281** (0%)	**0.320** (+7%)
	PoC	0.101	0.115	0.139	0.111	0.069
Shakespeare	FedAvg	0.54	0.549	0.522	0.540	0.538
	F3AST	**0.569** (+5%)	**0.555** (+1%)	**0.568** (+9%)	0.556 (+3%)	0.570 (+6%)
	FedAdam	0.536	0.541	0.520	**0.557**	0.520
	F3AST + Adam	0.549 (+2%)	0.551 (+2%)	0.566 (+9%)	**0.557** (0%)	0.551 (+6%)
	PoC	0.554	**0.555**	0.496	0.555	0.535

availability model, can work in conjuction with client unavailability. In PoC, the server at time t samples d clients from the available set C_t without replacement, choosing client k with probability p_k. The d clients receive the current model $\mathbf{w}_t$ and inform the server about their current loss $F_k(\overline{\mathbf{w}}^t)$. The server then selects for training the top M clients with the highest loss.

We do not compare our method with stateful techniques (Scaffold, FedDyn) since they are not applicable in the cross-device federated settings [10].

To evaluate performance of the algorithms, we compute the loss and accuracy using per-test-sample averages (the average is taken over individual data points).

A. Numerical Results

1) Accuracy: We first show the convergence of F3ASTon the three datasets with *Home-devices* availability model, a setting that fits Assumption 1 and is realistic in FL (note that the synthetic dataset satisfies all assumptions from Section III). Corroborating expectations of the impact of the de-biasing step introduced by F3AST, Fig. 2 shows that F3ASTachieves higher accuracy than FedAvg and PoCon all datasets (the corresponding loss plot can be found in Section C). Moreover, after the first 100 iterations, F3ASTstabilizes on Shakespeare and Synthetic datasets, and follows a more stable learning trajectory, illustrating its variance reduction advantage, unlike the two baselines that have high variability due to time-varying client availability. The sharp drop in Shakespeare is caused by sampling a client misaligned due to the heterogenous nature of the data; this has been reported in [41]. We show an average over three runs in Section C. We observe a similar behaviour on CIFAR100: F3ASTachieves almost a 200% improvement in the average accuracy over the last 100 rounds, and has a much more stable behaviour. The stagnation of FedAvg and PoCat higher loss models confirms that naive averaging introduces bias to the model, hindering convergence.

Table III shows the final accuracy of algorithms under diverse availability models defined in Section IV-2a (1000 rounds on CIFAR100 and 500 rounds on Shakespeare[4]). F3ASTeffectively improves the accuracy of FedAvg over both datasets and for all availability models. It also improves FedAdamfor all but the *Uneven* model where the accuracy remains the same although the loss value is lower (Table IV). F3ASTis particularly successful in difficult settings, e.g. *Scarce* and *Uneven*, where a small number of clients is available for training and where the client availability is inversely proportional to the amount of data clients hold – there, F3ASTis able to maintain performance similar to the setting where all clients are available. Meanwhile, the performance of both FedAvg and PoCdeteriorates. FedAdamis able to maintain

[4] Higher accuracy values on CIFAR100 could be obtained by running experiments for 10,000+ rounds

TABLE IV
Sample Loss of All Methods, and Relative Improvements of F3ASTOver FedAvg and F3AST+ Adam Over FedAdam, for Different Availability Models (Columns) on CIFAR100 (1000 Rounds) and Shakespeare (500 Rounds)

		Availability model				
		Always	Scarce	HomeDevice	Uneven	SmartPhones
CIFAR100	FedAvg	4.42	4.77	4.74	5.29	4.28
	F3AST	4.20 (-5%)	4.18 (-12%)	4.14 (-13%)	4.17 (-21%)	4.15 (-3%)
	FedAdam	3.74	3.65	3.50	3.67	3.55
	F3AST + Adam	**3.69** (-1%)	**3.61** (-1%)	**3.45** (-1%)	**3.66** (-0.5%)	**3.46** (-2%)
	PoC	4.90	4.84	4.63	4.86	5.42
Shakespeare	FedAvg	1.24	1.15	1.36	1.17	1.22
	F3AST	**1.10** (-11%)	**1.13** (-1%)	**1.10** (-19%)	1.13 (-3%)	**1.10** (-10%)
	FedAdam	1.27	1.23	1.40	1.12	1.36
	F3AST + Adam	1.18 (-8%)	1.19 (-3%)	1.11 (-21%)	**1.11** (-1%)	1.18 (-13%)
	PoC	1.13	**1.13**	1.74	1.13	1.24

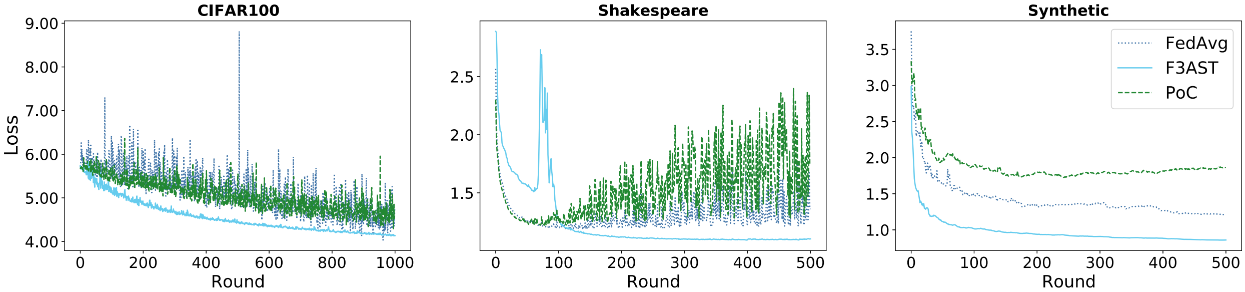

Fig. 3. Test per-sample loss of different algorithms over all data sets. In all cases F3ASTconverges to a model with smaller objective value. Furthermore, F3ASTstabilizes while FedAvg and PoCare not able to adapt to the time-varying environment.

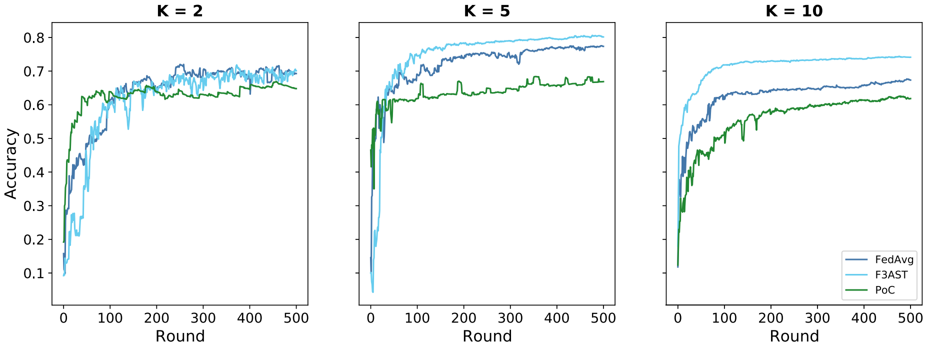

Fig. 4. Impact of varying communication constraint K in the Synthetic(1,1) dataset experiments. As the number of sampled clients increases, the gap between F3ASTand competing methods widens.

performance in the *Uneven* model where momentum may help with biased updates, but deteriorates in all other settings.

Finally, we note that F3AST achieves greater performance improvements in experiments on CIFAR100 than on Shakespeare. This is expected since, by design, F3ASTprovides more advantage in data heterogeneous settings where biased sampling may have a major detrimental effect on the performance/convergence of FL. Conversely, in the homogeneous settings, selecting one client more often than others does not affect the objective function because users are basically interchangeable. Both CIFAR100 and Shakespeare datasets are heterogeneous but there is considerably more heterogeneity in the federation of CIFAR100 where different users possess different, disjoint, categories in their local datasets. In contrast, Shakespeare is a next-word-prediction task where most clients have access to all the "categories" (words, in this case); while each client/character has a unique distribution over words, the common (English) language binds them together.

2) *Loss and Accuracy Values Under Independent Availability Model:* Fig. 3 shows that F3ASTexhibits a much more stable convergence for all data sets and achieves a smaller loss value.

B. *Varying the Communication Constraint*

Fig. 4 shows the test accuracy during training for the three algorithms (F3AST, FedAvg and PoC). We observe that F3ASTachieves equal or higher performance than competing methods across all communication levels. Note that

PoCstagnates at a similar accuracy in all cases; we believe this is due to the inherent bias in the algorithm due to top-k loss based sampling, as reported by the authors. It is possible that certain groups of clients are never selected by this policy. It is interesting to note that the gap between F3ASTand two baselines widens as the number of selected clients increases. Indeed, as the number of users grow, a configuration-dependent policy becomes much harder to facilitate since the number of possible selections grows exponentially with K. Nevertheless, the greedy nature of F3ASTallows it to keep selecting the set of users that maximizes marginal utility and achieves a balanced sampling rate under the availability model. The other two policies, however, do not track previous selections of users and thus may end up over-selecting available users rather than exploring the full pool of devices.

V. Conclusion

We presented F3AST, an algorithm for learning in federated systems that operate under communication constraints and service intermittently available clients. We demonstrated that the algorithm achieves accuracy superior to state-of-the-art federated learning techniques, and exhibits resilience in challenging system settings. Future work includes studies of the setting where the clients are grouped in clusters/classes, and exploring a wider range of communication constraints.

Appendices

Appendix A
Notation and Definitions

For clarity, frequently used symbols are summarized in Table V below .

The generalized FedAvg assumes the server sends to clients in S_t at time t an initial model $\overline{\mathbf{w}}^t$. For $t = 0, \ldots, T$; the clients locally initialize $\mathbf{w}_k^{(t,0)} \leftarrow \overline{\mathbf{w}}^{t-1}$ and take E steps of SGD producing the sequence $(\mathbf{w}_k^{(t,i)})_{i=0}^{E}$. Formally, let $\xi_k^{(t,i)}$ be the mini-batch for client k at time i in round t; for each client k, we can then define the local model $\mathbf{w}_k^{(t,i)}$ and local update $\mathbf{v}_k^t$ as

$$\mathbf{w}_k^{(t+1,i)} = \begin{cases} \overline{\mathbf{w}}^t & i = 0 \\ \mathbf{w}_k^{(t+1,i-1)} - \eta_{t+1,i} \nabla F_k(\mathbf{w}_k^{(t+1,i-1)}, \xi_k^{(t+1,i)}), & i \in [E] \end{cases}$$

$$\mathbf{v}_k^{t+1} = \mathbf{w}_k^{(t+1,E)} - \overline{\mathbf{w}}^t.$$

Here $\mathbf{w}_k^{(t,i)}$ tracks local models of client k at round t and iteration i, and $\mathbf{v}_k^t$ is the local update of client k at the end of round t. Following distributed optimization standard techniques [12], [35], we define the sequences

$$\begin{aligned} \overline{\mathbf{v}}^{t+1} &= \sum_{k=1}^{N} p_k \mathbf{v}_k^{t+1}, \\ \Delta^{t+1} &= \sum_{i \in S_t} \frac{p_i}{r_i} \mathbf{v}_k^{t+1}, \\ \overline{\mathbf{w}}^{t+1} &= \overline{\mathbf{w}}^t + \Delta^{t+1}, \\ \overline{\mathbf{z}}^{t+1} &= \overline{\mathbf{w}}^t + \overline{\mathbf{v}}^{t+1}. \end{aligned} \tag{9}$$

Appendix B
Proofs

This section provides proofs of the lemmas and theorems omitted from the main document.

A. Proof of Theorem III.5

Proof of Theorem III.5 follows standard optimization proofs [5], [12], [35]. Our technical contribution comes from using a sampling policy with arbitrary sampling rate $\mathbf{r}$, showing global update Δ^{t+1} is unbiased (Lemma B.1), and computing the incurred variance of such sampling policy and aggregation step (Lemma III.4). This last step is of particular interest and challenging due to the unknown system configuration, and the importance sampling multiplicative terms.

Lemma B.1 (Unbiased update): Suppose Assumption 1 holds. Let $\mathbf{r} \in \mathcal{R}$ be an achievable sampling rate, and $\boldsymbol{f}^r$ be a state-dependent static policy achieving rate $\mathbf{r}$. Fix $\overline{\mathbf{w}}^t \in \mathbb{R}^p$. Let $(\mathbf{v}_k^{t+1})_{k=1}^N$ denote updates of clients starting from model $\overline{\mathbf{w}}^t$. Let $\overline{\mathbf{v}}^{t+1} = \sum_{k=1}^N p_k \mathbf{v}_k^{t+1}$, let S be the client set selected by policy $\boldsymbol{f}^r$ at time t and $\Delta^{t+1} = \sum_{k \in S} \frac{p_k}{r_k} \mathbf{v}_k^{t+1}$. Then $\mathbb{E}_S[\Delta^{t+1}] = \overline{\mathbf{v}}^{t+1}$.

Proof: Recall that at any given time t we pick $S \in C_t$ for some $C_t \in \mathcal{C}$ according to $f_{C_t S}^r$. Using the definition of π and $\boldsymbol{f}^r$,

$$\mathbb{E}_S\left[\Delta^{t+1}\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k \in S} \frac{p_k}{r_k} \mathbf{v}_k^{t+1} \middle| C\right]\right] \tag{10}$$

$$= \sum_{C \in \mathcal{C}} \pi(C) \sum_{S \in C} f_{CS}^r \sum_{k \in S} \frac{p_k}{r_k} \mathbf{v}_k^{t+1} \tag{11}$$

$$= \sum_{C \in \mathcal{C}} \pi(C) \sum_{S \in C} f_{CS}^r \sum_{k=1}^{N} \frac{p_k}{r_k} \mathbf{v}_k^{t+1} \cdot \mathbb{1}_{\{k \in S\}}, \tag{12}$$

where the last step replaces the sum over S by the sum over all clients but adding the indicator function over S. Reorganizing,

$$\mathbb{E}_S\left[\Delta^{t+1}\right] = \sum_{k=1}^{N} \frac{p_k}{r_k} \mathbf{v}_k^{t+1} \left(\sum_{C \in \mathcal{C}} \pi(C) \sum_{S \in C} f_{CS}^r \mathbb{1}_{\{k \in S\}}\right). \tag{13}$$

Here the term in parenthesis is, by definition (Eq. (2)), r_k, then

$$= \sum_{k=1}^{N} \frac{p_k}{r_k} \mathbf{v}_k^{t+1} r_k = \sum_{k=1}^{N} p_j \mathbf{v}_{t+1}^{j} = \overline{\mathbf{v}}^{t+1}.$$

■

We proceed by introducing a key lemma derived in [12], characterizing convergence for the full client participation case, and then utilize it to prove Theorem III.5.

Notice that the learning rate depends on round $t, t = 0, \ldots, T$ and epoch $i, i = 1, \ldots, E$.

Lemma B.2:

$$\mathbb{E}[\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2] \le (1 - \eta_{(t,E)}^2 \mu) \mathbb{E}[\|\overline{\mathbf{w}}^t - \mathbf{w}^*\|^2] + \eta_{(t,E)}^2 \mathrm{Var}_1, \tag{14}$$

TABLE V
FREQUENTLY USED SYMBOLS

Symbol	Definition
$\mathcal{U}$	Set of all users
$N = \lvert\mathcal{U}\rvert$	Total number of clients
K_t	Number of clients sampled at round t
T	Total number of rounds
m_t	Number of available clients at time t
$(\mathbf{A}_t)_t$	availability stochastic process
$(\mathbf{C}_t)_t$	feasible client sets stochastic process
S_t	Set of clients participating at round t
$\pi(\cdot)$	Stationary distribution of availability and communication constraint process.
$\mathbf{r}$	Client sampling rate
$\boldsymbol{f}^{\mathbf{r}}$	Configuration-dependent client sampling policy
Δ^{t+1}	Pseudo-gradient for server optimizer: aggregates updates of participating clients at round t
$\overline{\mathbf{w}}^t$	Global model at beginning of round t
$\mathbf{w}_k^{t+1}$	Local model at the end of round t at client k
$\mathbf{v}_k^{t+1}$	Local update at the end of round t at client k
$\overline{\mathbf{v}}^{t+1} = \mathbb{E}_{k\sim\mathcal{P}}\left[\mathbf{v}^k_{t+1}\right] = \sum_{k=1}^N p_k \mathbf{v}^k_{t+1}$	Expected global update at the end of round t under desired distripution $\mathcal{P}$
$\overline{\mathbf{z}}^{t+1} = \overline{\mathbf{w}}^t + \overline{\mathbf{v}}^{t+1}$	Desired global model at the end of round t

where

$$\mathrm{Var}_1 = \sum_{k=1}^N p_k^2\sigma_k^2 + 6L\Gamma + 8(E-1)^2G^2.$$

Proof: This result follows from the first part of Theorem 1 in [12], showing the convergence of FEDAVG with full client participation (i.e., no client sampling).

Theorem (Theorem III.5): Instate the settings of Lemma III.4. Let $\mathbf{w}^*$ denote the solution to the optimization problem (1), and $L, \mu = O(1)$. Define $\gamma = \max\{8\frac{L}{\mu}, E\}$, and assume learning rate $\eta_{(t,i)} = \frac{2}{\mu(\gamma+(tE+i))}$. Then by setting CLIENTOPT to SGD and SERVEROPT$(\overline{\mathbf{w}}^t, \Delta^{t+1}) = \overline{\mathbf{w}}^t + \Delta^{t+1}$, the model $\overline{\mathbf{w}}^T$ produced by Algorithm 1 with policy $\boldsymbol{f}^{\mathbf{r}}$ satisfies

$$\mathbb{E}[F(\overline{\mathbf{w}}^T)] - F^* = O\left(\frac{1}{TE+\gamma}\left(\|\overline{\mathbf{w}}^1 - \mathbf{w}^*\|^2 \right.\right.$$
$$+ \sum_{k=1}^N p_k^2\sigma_k^2 + 6\Gamma + 8(E-1)^2G^2$$
$$\left.\left. + \sigma_T^2(\boldsymbol{f}^{\mathbf{r}})\right)\right),$$

where $\Gamma = F^* - \sum_{k=1}^N p_k F_k^*$ denotes the local-global objective gap[5], F^* and F_k^* are the minimum values of F and F_k, respectively, and $\sigma_T^2(\boldsymbol{f}^{\mathbf{r}})$ (Lemma III.4) captures the variance induced by client sampling.

a) A brief outline of the upcoming proof: We start by expanding $\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2$, a term that measures the distance to the optimum, and bound it by the term characterizing convergence of the full participation scheme [7] plus an additional variance term emerging due to client sampling (computed in Lemma B.5). We then invoke a standard inductive argument to express $\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2$ in terms of $\|\overline{\mathbf{w}}^1 - \mathbf{w}^*\|^2$ and, noting smoothness, finally bound $\mathbb{E}[F(\mathbf{w}_T)] - F^*$.

Proof:

$$\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2 = \|\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{z}}^{t+1} + \overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\|^2$$
$$= \underbrace{\|\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{z}}^{t+1}\|^2}_{A_1} + \underbrace{\|\overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\|^2}_{A_2}$$
$$+ \underbrace{2\langle\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{z}}^{t+1}, \overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\rangle}_{A_3}. \qquad (15)$$

Based on Lemma B.1, we know Δ^{t+1} is unbiased, thus $\mathbb{E}[\Delta^{t+1}] = \overline{\mathbf{v}}^{t+1}$.

We use this fact to prove that $A_3 = 0$ as follows:

$$\mathbb{E}[\langle\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{z}}^{t+1}, \overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\rangle] =$$
$$\mathbb{E}\left[\langle\overline{\mathbf{w}}^t + \Delta^{t+1} - \overline{\mathbf{w}}^t - \overline{\mathbf{v}}^{t+1}, \overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\rangle\right]$$
$$= \mathbb{E}\left[\langle\Delta^{t+1} - \overline{\mathbf{v}}^{t+1}, \overline{\mathbf{z}}^{t+1} - \mathbf{w}^*\rangle\right]$$
$$= 0$$

Now, A_1 can be bounded using Lemma III.4, since $\|\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{z}}^{t+1}\|^2 = \|(\overline{\mathbf{w}}^t + \Delta^{t+1}) - (\overline{\mathbf{w}}^t + \overline{\mathbf{v}}^{t+1})\|^2 = \|\Delta^{t+1} - \overline{\mathbf{v}}^{t+1}\|^2$.

Define

$$\mathrm{Var}_1 = \sum_{k=1}^N p_k^2\sigma_k^2 + 6\Gamma + 8(E-1)^2G^2,$$

$$\mathrm{Var}_2 := \sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) = \frac{1}{\eta_{(t,E)}^2}\mathbb{E}\left[||\Delta^{t+1} - \overline{\mathbf{v}}^{t+1}\|^2\right].$$

Then by replacing Lemmas B.2 and B.5 in (15) we have that

$$\mathbb{E}\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2 \le (1 - \eta_{(t,E)}\mu)\mathbb{E}\|\overline{\mathbf{w}}^t - \mathbf{w}^*\|^2$$
$$+ \eta_{(t,E)}^2(\mathrm{Var}_1 + \mathrm{Var}_2).$$

Thanks to Lemma B.1, we find a similar expression to the one in [12], but with different constants coming from different client sampling variance in Lemma III.4. The rest of the proof then follows standard techniques, e.g., see [12]. We repeat those steps for the sake of completeness.

Let $\beta > \frac{1}{\mu}$, $\gamma > 0$, and define $\eta_{(t,i)} = \frac{\beta}{(t-1)E+i+\gamma}$ such that $\eta_{(1,1)} < \min\{\frac{1}{\mu}, \frac{1}{4L}\}$, and $\eta_{(t,1)} \le 2\eta_{(t,E)}$.

[5] Local-global objective gap quantifies data heterogeneity: for i.i.d. data, $\Gamma \to 0$ as the number of samples grows, while a large Γ indicates a high degree of heterogeneity [11], [12].

As a standard technique, we show by induction in t that $\mathbb{E}\|\overline{\mathbf{w}}^t - \mathbf{w}^*\|^2 \leq \frac{v}{\gamma + tE}$, where

$$v = \max\{\frac{\beta^2(\mathrm{Var}_1 + \mathrm{Var}_2)}{\beta\mu - 1}, (\gamma+1)\|\overline{\mathbf{w}}^1 - \mathbf{w}^*\|^2\}.$$

This holds for $t = 1$ trivially, from the definition of v. Now assume the claim holds for t; starting from the above equation, we have that

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\|^2 &\leq (1 - \eta_{(t,E)}\mu)\mathbb{E}\|\overline{\mathbf{w}}^t - \mathbf{w}^*\|^2 \\
&\quad + \eta_{(t,E)}^2(\mathrm{Var}_1 + \mathrm{Var}_2) \\
&\leq \left(1 - \frac{\beta\mu}{tE+\gamma}\right)\frac{v}{tE+\gamma} \qquad (16)\\
&\quad + \frac{\beta^2(\mathrm{Var}_1 + \mathrm{Var}_2)}{(tE+\gamma)^2} \\
&= \frac{tE+\gamma-E}{(tE+\gamma)^2}v \qquad (17)\\
&\quad + \left[\frac{\beta^2(\mathrm{Var}_1 + \mathrm{Var}_2)}{(tE+\gamma)^2} - \frac{\beta\mu - E}{(tE+\gamma)^2}v\right] \\
&\leq \frac{v}{tE+\gamma+E} \\
&= \frac{v}{(t+1)E+\gamma}. \qquad (18)
\end{aligned}$$

where (16) follows by induction step and definition of $\eta_{(t,E)}$, (17) from adding and substracting $\frac{Ev}{(tE+\gamma)^2}$, and the last step by noticing that

$$\begin{aligned}
\frac{tE+\gamma-E}{(tE+\gamma)^2} &= \frac{(tE+\gamma-E)(tE+\gamma+E)}{(tE+\gamma)^2(tE+\gamma+E)} \\
&\leq \frac{1}{tE+\gamma+E}.
\end{aligned}$$

Finally, by smoothness of F,

$$\mathbb{E}[F(\overline{\mathbf{w}}^t)] - F^* \leq \frac{L}{2}\mathbb{E}\|\overline{\mathbf{w}}^t - \mathbf{w}^*\|^2 \leq \frac{v}{\gamma + tE}. \qquad (19)$$

Setting $\beta = \frac{2}{\mu}$, $\kappa = \frac{L}{\mu}$ and $\gamma = \max\{8\kappa, E\} - 1$, and using Lemma B.5 to compute Var_2, we obtain the desired result. ■

B. Proof of Lemma III.4: Bounded Client Sampling Variance

For clarity of the proof of Lemma III.4, we first introduce the following lemma on the inner product of local models.

Lemma B.3: At round t for any pair of clients i and j,

$$\mathbb{E}\left[\langle \mathbf{v}_i^t, \mathbf{v}_j^t\rangle\right] \leq 4E^2G^2\eta_{(t,E)}^2, \qquad (20)$$

where the expectation is taken over the samples in local SGD steps.

Proof: Recall that $\mathbf{v}_i^t$ represents the user i's update after training locally for E epochs starting with model $\overline{\mathbf{w}}^{t-1}$. Then,

$$\begin{aligned}
\|\mathbf{v}_i^t\|^2 = \|\mathbf{w}_i^{t,E} - \mathbf{w}_i^{t,0}\|^2 &= \left\|\sum_{\ell=0}^{E-1}\eta_{(t,\ell)}\nabla f_k(\mathbf{w}_i^{t,\ell}, \xi_\ell)\right\|^2 \\
&\leq E\sum_{\ell=0}^{E-1}\eta_{t,0}^2\|\nabla f_k(\mathbf{w}_i^{t,\ell})\|^2 \leq 4\eta_{(t,E)}^2E^2G^2,
\end{aligned}$$

where we used the Jensen inequality and the fact that $\eta_{(t,\ell)}$ is decreasing (i.e., $\eta_{(t,0)} \leq 2\eta_{(t,\ell)}$ for $\ell \leq E$). Now, since $\langle \mathbf{v}_i^t, \mathbf{v}_j^t\rangle \leq \max_k \|\mathbf{v}_k^t\|^2$, the result follows. ■

Next, we define the random vector $X^{\mathbf{r}} \in \{0,1\}^N$ that indicates which clients are selected, and specify its moments.

Lemma B.4: Let $X^{\mathbf{r}} \in \{0,1\}^N$ be such that its i^{th} component takes on value $X_i^{\mathbf{r}} = 1$ if client i is selected at time t, and 0 otherwise. Let Σ denote the covariance matrix of $X^{\mathbf{r}}$. Then $\mathbb{E}[X] = \mathbf{r}$ and $\mathrm{Var}(X_i) = r_i(1 - r_i)$.

Proof: This follows trivially from the observation that X_i is a Bernoulli random variable with parameter r_i. ■

Lemma B.5: Let $\mathbf{r} \in \mathcal{R}^+$ be an achievable sampling rate, $\boldsymbol{f}^{\mathbf{r}}$ denote a static configuration-dependent sampling policy achieving $\mathbf{r}$, and $X^{\mathbf{r}}$ the corresponding selection random vector with covariance $\boldsymbol{\Sigma}$. Then for $t = 1, \ldots, T$,

$$\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) := \frac{1}{\eta_{(t,E)}^2}\mathbb{E}\left[\|\Delta^t - \overline{\mathbf{v}}^t\|^2\right] = \frac{1}{\eta_{(t,E)}^2}\mathbf{Tr}(\mathbf{Y}_t\mathbf{Y}_t^T\boldsymbol{\Sigma}), \qquad (21)$$

where vector $\frac{p_k}{r_k}\mathbf{v}_k^t$ is the k^{th} row of matrix $\mathbf{Y}_t \in \mathbb{R}^{N\times p}$.

Proof: Below $\|\cdot\|$ denotes the ℓ_2-norm. Using the variance formula and that $\mathbb{E}[\Delta^t] = \overline{\mathbf{v}}^t$ by Lemma B.1,

$$\mathbb{E}_S\left[\|\Delta^t - \overline{\mathbf{v}}^t\|^2\right] = \mathbb{E}\left[\|\Delta^t\|^2\right] - \|\overline{\mathbf{v}}\|^2. \qquad (22)$$

Let us focus on the first term:

$$\begin{aligned}
\mathbb{E}\left[\|\Delta^t\|^2\right] &= \sum_{C\in\mathcal{C}}\pi(C)\sum_{S\in C}f_{CS}\left\|\sum_{k\in S}\frac{p_k}{r_k}\mathbf{v}_k^t\right\|^2 \\
&= \sum_{C\in\mathcal{C}}\pi(C)\sum_{S\in C}f_{CS}\sum_{i,j\in S}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i, \mathbf{v}_j\rangle \\
&= \sum_{i,j=1}^{N}\sum_{C\in\mathcal{C}}\pi(C)\sum_{S\in C}f_{CS}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i, \mathbf{v}_j\rangle\mathbb{1}_{i\in S}\mathbb{1}_{j\in S} \\
&= \sum_{i,j=1}^{N}\sum_{C\in\mathcal{C}}\pi(C)\sum_{S\in C}f_{CS}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i, \mathbf{v}_j\rangle X_iX_j.
\end{aligned}$$

From the defition of $\mathbf{Y}_t$ and by reorganizing,

$$\mathbb{E}\left[\|\Delta^t\|^2\right] = \mathbb{E}\left[X^T\mathbf{Y}_t\mathbf{Y}_t^TX\right].$$

Introducing $\mathbf{B} = \mathbf{Y}_t\mathbf{Y}_t^T$,

$$\begin{aligned}
\mathbb{E}\left[\|\Delta^t\|^2\right] &= \mathbb{E}\left[\sum_{i,j}b_{ij}X_iX_j\right] = \sum_{i,j}b_{ij}\mathbb{E}\left[X_iX_j\right] \\
&= \sum_{i,j}b_{ij}(\boldsymbol{\Sigma}_{ij} + r_ir_j) \qquad (23)\\
&= \sum_i[\mathbf{B}\boldsymbol{\Sigma}]_{ii} + \mathbf{r}^T\mathbf{B}\mathbf{r} \\
&= \mathbf{Tr}(\mathbf{Y_tY_t^T\Sigma}) + \|\overline{\mathbf{v}}^t\|^2, \qquad (24)
\end{aligned}$$

where Eq. 23 follows by the covariance formula and Eq. 24 follows due to cancellation in denominator of $\mathbf{B} = \mathbf{Y}_t\mathbf{Y}_t^T$. After combining this with (22), the term $\|\overline{\mathbf{v}}^t\|^2$ cancels and we obtain the desired result. ■

Lemma (Lemma III.4) Suppose Assumptions 1-4 hold. Let $\mathbf{r} \in \mathcal{R}$ be an achievable sampling rate under the system configuration determined by distribution π, and $\boldsymbol{f}^{\mathbf{r}}$ denote a static configuration-dependent sampling policy achieving $\mathbf{r}$.

Define the client sampling variance $\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) := \frac{1}{\eta_t^2}\mathbb{E}_{S_t}[\|\Delta^t - \overline{\mathbf{v}}^t\|^2]$ where $\overline{\mathbf{v}}^t = \sum_{k=1}^N p_k \mathbf{v}_t^k$ is the update at time t with full client participation. Then

$$\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) \leq 4E^2G^2\left(\sum_{k=1}^N \frac{p_k}{r_k} - 1\right). \tag{25}$$

Furthermore, if client availabilities are uncorrelated or negatively correlated, then there exists a policy $\boldsymbol{f}^r$ such that

$$\sigma_t^2(\boldsymbol{f}^{\mathbf{r}}) \leq 4E^2G^2\left(\sum_{k=1}^N \frac{p_k^2}{r_k} + \sum_{k=1}^N p_k^2\right). \tag{26}$$

Proof: For the first part, taking expectation over the independent local SGD sampling and over the random set of clients S,

$$\begin{aligned}
\mathbb{E}\left[\|\Delta^t\|^2\right] &= \mathbb{E}\left[\|\sum_{k\in S}\frac{p_k}{r_k}\mathbf{v}_k^t]\|^2\right] \\
&= \mathbb{E}\left[\sum_{i,j\in S}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i^t, \mathbf{v}_j^t\rangle\right] \\
&= \sum_{i,j=1}^N \frac{p_ipj}{r_ir_j}\mathbb{E}\left[\langle\mathbf{v}_i^t, \mathbf{v}_j^t\rangle\right]P(i,j\in S) \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_{i=1}^N \frac{p_i^2}{r_i^2}P(i\in S)\right. \\
&\quad \left. + \sum_{i=1}^N\sum_{j=1,j\neq i}^N \frac{p_ip_j}{r_ir_j}P(i,j\in S)\right).
\end{aligned}$$

Given that $P(i,j\in S) \leq P(i\in S)$ and that $P(i\in S) = r_i$,

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta^t\|^2\right] \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_{i=1}^N\frac{p_i^2}{r_i} + \sum_{i=1}^N\sum_{j=1,j\neq i}^N\frac{p_ip_j}{r_ir_j}P(j\in S)\right) \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_{i=1}^N\frac{p_i^2}{r_i} + \sum_{i=1}^N\frac{p_i}{r_i}\sum_{j=1,j\neq i}^N p_j\right) \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_{i=1}^N\frac{p_i^2}{r_i} + \sum_{i=1}^N\frac{p_i}{r_i}(1-p_i)\right) \\
&\leq 4E^2G^2\eta_{(t,E)}^2\sum_{i=1}^N\frac{p_i}{r_i}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\Delta^t - \overline{\mathbf{v}}^t\|^2\right] &= \mathbb{E}\left[\|\Delta^t\|^2\right] - \|\overline{\mathbf{v}}^t\|^2 \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_{i=1}^N\frac{p_i}{r_i} - 1\right),
\end{aligned}$$

and the result follows.

For the second part of the lemma, consider a policy $\boldsymbol{f}$ with rate $\mathbf{r}$ that at time t selects S as

$$S_t \in \arg\max_{S\in C_t} -\nabla H(\mathbf{r})\cdot\mathbb{1}_S.$$

Let u_k denote the k-th largest utility value, where the individual utilities are defined by vector $-\nabla H(\mathbf{r})$. W.l.o.g. assume $u_i < u_j$ if $i < j$, and let A_k be the (random) number of users i with the utility less than u_k. Let K be a bound on the set size S. Let i, j be two users, $i < j$; then $u_i < u_j$. Now, since i, j are uncorrelated or negatively correlated, $P(i, j \text{ are available}) \leq P(j \text{ is available})P(i \text{ is available})$. Therefore,

$$\begin{aligned}
P(j\in S|i\in S) &= P(j \text{ is available})P(A_k - 1 < K) \\
&\leq P(j \text{ is available})P(A_k < K) = P(j\in S).
\end{aligned}$$

Note that from the definition of conditional probability, it follows that the sampling is also uncorrelated since

$$\begin{aligned}
P(i,j\in S) &= P(i\in S)P(j\in S|i\in S) = P(i\in S)P(j\in S) \\
&= \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\
&= r_ir_j.
\end{aligned}$$

Therefore, we have that $\Sigma_{ij} = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = P(i,j\in S) - r_ir_j \leq 0$. From Eq. 23, we have that

$$\mathbf{Tr}(\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\boldsymbol{\Sigma}) = \sum_{i,j}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i^t, \mathbf{v}_j^t\rangle\boldsymbol{\Sigma}_{ij};$$

since $\boldsymbol{\Sigma}_{ij} \leq 0$ for $i \neq j$,

$$\begin{aligned}
\mathbf{Tr}(\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\boldsymbol{\Sigma}) &= \sum_{i,j}\frac{p_ip_j}{r_ir_j}\langle\mathbf{v}_i^t, \mathbf{v}_j^t\rangle\boldsymbol{\Sigma}_{ij} \\
&\leq 4E^2G^2\eta_{(t,E)}^2\sum_i\frac{p_i^2}{r_i^2}\Sigma_{ii} + \sum_{i\neq j}\frac{p_ip_j}{r_ir_j}\Sigma_{ij} \\
&\leq 4E^2G^2\eta_{(t,E)}^2\sum_i\frac{p_i^2}{r_i^2}(r_i(1-r_i)) \\
&\leq 4E^2G^2\eta_{(t,E)}^2\left(\sum_i\frac{p_i^2}{r_i} + \sum_i p_i^2\right),
\end{aligned}$$

where the first inequality follows from Lemma B.3 and breaking the sum on diagonal and non diagonal terms. The second inequality follows by dropping negative terms and replacing the variance value $\Sigma_{ii} = r_i(1-r_i)$ for X_i (Lemma B.4), while the last one follows simply by expanding the previous term. ■

C. Proof of Theorem III.3

Theorem (Theorem III.3): Let $\mathbf{r}^\beta(t)$ be defined by Algorithm 1, following equations (4) and (5). Let $V \subset \mathbb{R}_+^N$ be a bounded set, $\epsilon > 0$, and let $\mathbf{r}^*$ denote the minimizer of the variance function $H(\mathbf{r}) = \sum_{k=1}^{n} \frac{p_k^2}{r_k}$ over $\mathcal{R}$. Then for $T > 0$, depending on ϵ and V,

$$\lim_{\beta \downarrow 0} \sup_{\mathbf{r}^\beta(0) \in V, t > T/\beta} P[\|r^\beta(t) - r^*\| > \epsilon] = 0.$$

Proof: Our proof follows the stochastic approximation analysis in [43] that establishes an attraction property for Fluid Sample Paths (FSPs), which are limiting trajectories of the generalized versions of the processes considered in our manuscript. The idea behind the proof is that as t grows, we can study the limiting trajectories (i.e., FSPs) $x = (x(t), t \geq 0)$ of the process $r(t/\beta)$. Since $-H(\mathbf{r})$ is a convex function, it follows from Theorem 4 in [43] that for an arbitrary initial state $x(0)$, $x(t) \to \mathbf{r}^*$ as $t \to \infty$. More concretely, from Theorem 3 in [43] it follows that as $\beta \downarrow 0$, a limit of sequence $\{\mathbf{r}^\beta\}$ considered in Section III of our paper is a process with sample paths being FSPs x with probability 1.

Assume $V \subset [0, a]^N$ such that $a > \bar{\mu}$, where $\bar{\mu} = \max_{C \in \mathcal{C}, S \in S} \frac{1}{|S|}$. Let $\epsilon > 0$ and $\delta > 0$. By the above result on FSPs (i.e., by Theorem 4 in [43]), we can find T large enough such that $\|x(t) - \mathbf{r}^*\| \leq \epsilon$ uniformly for t in the interval $[T, T + \delta]$. Combining this result with the continuous mapping theorem [44], we obtain

$$\lim_{\beta \to 0} \sup_{r^\beta \in V} P\left(\sup_{t \in [T, T+\delta]} \|x(t) - \mathbf{r}^*\| > \epsilon \right) = 0.$$

Also, notice that for all t, $r_i^\beta(t) \leq \max\{\bar{\mu}, \mathbf{r}_i^\beta(0)\}$ element-wise, where $\bar{\mu} = \max_{C \in \mathcal{C}, S \in S} \frac{1}{|S|}$ by construction. Then for any $\tau \geq 0$ we can re-start the process, implying

$$\lim_{\beta \to 0} \sup_{r^\beta \in V} \sup_{\tau \geq 0} P\left(\sup_{t \in [\tau + T, \tau + T + \delta]} \|x(t) - \mathbf{r}^*\| > \epsilon \right) = 0$$

and thus establishing the desired result.

Appendix C
Experiments Details

a) Hyperparameter tuning: We set the learning rate on Shakespeare and CIFAR100 according to the optimal values found in [5]. For the synthetic dataset we use the learning rate tuned in [7], $\eta = 0.01$. For Shakespeare we use mini-batches of size 4, and mini-batches of size 20 for the remaining datasets. Following literature, in all the experiments we use $\beta = O(1/T) = 0.001$.

b) Machines: We ran our experiments on AMD Vega 20 (ROCm) cards. One rounds of training in Fig. 1 require 8 GPU seconds for CIFAR100, 67 GPU seconds for Shakespeare, and 0.57 GPU seconds for Synthetic(1,1).

c) Synthetic dataset: We generate this data by taking 10^4 samples $X_i \in \mathbb{R}^{100} \sim \mathcal{N}(0, I_{100})$. Moreover, we generate $\beta \sim \mathcal{N}(0, I_{100})$ and, finally, set labels $y_i = round(X_i^T \beta)$. The samples are split evenly among 100 clients.

d) Skakespeare: Each client's dataset is restricted to have at most 128 sentences, and is split into training and validation sets. Following the previous work with this dataset [5], we use a build vocabulary with 86 characters contained in the text, and 4 characters representing padding, out-of-vocabulary, beginning, and end of line tokens. We use padding and truncation to enforce 20 word sentences, and represent them with index sequences corresponding to the vocabulary words, out of vocabulary words, beginning and end of sentences.

A. Models

We train a recursive neural network for the next character prediction that first embeds characters into an 8-dimensional space, followed by 2 LSTMs and finally a dense layer. ResNet-18 architecture can be found in [45], where we replace batch normalization by group normalization [46] as in [5].

B. Availability Models

For the *Home-devices* model $t_k \sim lognormal(0, 0.5)$, while for the *Smartphones* $t_k \sim lognormal(0, 0.25)$. The sine wave is defined by $f(t) = 0.4\sin(t) + 0.5$ and we sample at times $t = \frac{2\pi j}{24}$ for $j = 1, \ldots, 24$.

Acknowledgement

This work was done while Mónica Ribero was affiliated with University of Texas at Austin, Austin, USA.

References

[1] K. Bonawitz et al., "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, 2019, pp. 374–388.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[4] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.

[5] S. Reddi et al., "Adaptive federated optimization," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/pdf?id=LkFG3lB13U5

[6] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *Proc. Int. Conf. Learn. Representations*, 2021.

[7] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2018, pp. 429–450.

[8] H. Eichner, T. Koren, B. McMahan, N. Srebro, and K. Talwar, "Semi-cyclic stochastic gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1764–1773.

[9] J. Wang et al., "A field guide to federated optimization," 2021, *arXiv:2107.06917*.

[10] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, pp. 1–210, 2019.

[11] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," 2020, *arXiv:2010.01243*.

[12] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Representations*, 2020.

[13] M. Ribero and H. Vikalo, "Communication-efficient federated learning via optimal client sampling," 2020, *arXiv:2007.15197*.

[14] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate Newton-type method," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1000–1008.

[15] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4424–4434.

[16] H. Tang, X. Lian, T. Zhang, and J. Liu, "DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6155–6165.

[17] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. Workshop Private Multi-Party Mach. Learn.*, 2016.

[18] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3329–3337.

[19] J. Konečný and P. Richtárik, "Randomized distributed mean estimation: Accuracy vs. communication," *Front. Appl. Math. Statist.*, vol. 4, 2018, Art. no. 62.

[20] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.

[21] S. Horvath, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtarik, "Natural compression for distributed deep learning," in *Proc. Math. Sci. Mach. Learn.*, 2022, pp. 129–141.

[22] W. Xia, T. Q. S. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit-based client scheduling for federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7108–7123, Nov. 2020.

[23] A. Hard et al., "Federated learning for mobile keyboard prediction," 2018, *arXiv:1811.03604*.

[24] T. Yang et al., "Applied federated learning: Improving Google Keyboard query suggestions," 2018, *arXiv:1812.02903*.

[25] K. Hsieh et al., "GAIA: Geo-distributed machine learning approaching {LAN } speeds," in *Proc. 14th {USENIX} Symp. Networked Syst. Des. Implementation*, 2017, pp. 629–647.

[26] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5050–5060.

[27] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized stochastic optimization," 2019, *arXiv:1910.14280*.

[28] E. Rizk, S. Vlaski, and A. H. Sayed, "Federated learning under importance sampling," *IEEE Trans. Signal Process.*, vol. 70, pp. 5381–5396, 2022.

[29] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–9.

[30] Y. Chen, Y. Ning, M. Slawski, and H. Rangwala, "Asynchronous online federated learning for edge devices with non-IID data," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 15–24.

[31] J. Nguyen et al., "Federated learning with buffered asynchronous aggregation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3581–3607.

[32] M. Rabi, G. V. Moustakides, and J. S. Baras, "Adaptive sampling for linear state estimation," *SIAM J. Control Optim.*, vol. 50, no. 2, pp. 672–702, 2012.

[33] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4615–4625.

[34] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3321–3363, 2013.

[35] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Int. Conf. Learn. Representations*, 2019.

[36] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4452–4463.

[37] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9324–9333.

[38] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, vol. 107. Ann Arbor, MI, USA: Amer. Math. Soc., 2017.

[39] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4387–4398.

[40] TFF The Authors, "TensorFlow federated," 2019. [Online]. Available: https://www.tensorflow.org/federated

[41] Z. Charles, Z. Garrett, Z. Huo, S. Shmulyian, and V. Smith, "On large-cohort training for federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 20461–20475.

[42] S. Alouf, E. Altman, and P. Nain, "Optimal online estimation of the size of a dynamic multicast group," in *Proc. IEEE 21st Annu. Joint Conf. Comput. Commun. Soc.*, 2002, vol. 2, pp. 1109–1118.

[43] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Res.*, vol. 53, no. 1, pp. 12–25, 2005.

[44] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 2013.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[46] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

Mónica Ribero received the B.Sc. degree in mathematics from the Universidad de los Andes, Bogotá, Colombia, in 2015, and the Ph.D. degree in electrical and computer engineering from the University of Texas, Austin, TX, USA, in 2022. She is currently working on federated learning under privacy and communication constraints with the University of Texas. She joined Google Research New York, NY, USA, as a Research Scientist. She held Research Internship Positions, Bell Laboratories, Murray Hill, NJ, USA, in 2008, CognitiveScale, Austin, TX, in 2019, and Google Research in 2020.

Haris Vikalo (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Zagreb, Zagreb, Croatia, in 1995, the M.S. degree in electrical engineering from Lehigh University, Bethlehem, PA, USA, in 1997, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2003. He has held a short-term appointment with Bell Laboratories, Murray Hill, NJ, USA, in the summer of 1999. From January 2003 to July 2003, he was a Postdoctoral Researcher, and from July 2003 to August 2007, he was an Associate Scientist with the California Institute of Technology, Pasadena, CA, USA. Since September 2007, he has been with the Department of Electrical and Computer Engineering, the University of Texas, Austin, TX, USA. His research interests include signal processing, machine learning, communications, and bioinformatics. Prof. Vikalo was the recipient of the 2009 National Science Foundation Career Award.

Gustavo de Veciana (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, CA, USA, in 1993. He is currently a Professor and Associate Chair with the Department of Electrical and Computer Engineering. From 2003–2007, he was the Director and Associate Director with the Wireless Networking and Communications Group. His research interests include the design, analysis and control networks, information theory and applied probability, measurement, modeling and performance evaluation, wireless and sensor networks, architectures and algorithms to design reliable computing, and networked systems. Dr. de Veciana is currently the Editor at large of IEEE/ACM TRANSACTIONS ON NETWORKING. He was the recipient of the Cockrell Family Regents Chair in engineering at University of Texas, Austin, TX, USA, an NSF CAREER Award 1996, and a Co-recipient of seven Best Paper awards including the 2021 IEEE Communication Society W. Bennett Prize. In 2009, he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He currently on the Board of Trustees of IMDEA Networks Madrid.