Federated Learning at Scale: Addressing Client Intermittency and Resource Constraints

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

***Abstract*—In federated learning systems, a server coordinates the training of machine learning models on data distributed across a number of participating client devices. In each round of training, the server selects a subset of devices to perform model updates and, in turn, aggregates those updates before proceeding to the next round of training. Most state-of-the-art federated learning algorithms assume that the clients are always available to perform training – an assumption readily violated in many practical settings where client availability is intermittent or even transient; moreover, in systems where the server samples from an exceedingly large number of clients, a client will likely participate in at most one round of training. This can lead to biasing the learned global model towards client groups endowed with more resources. In this paper, we consider systems where the clients are naturally grouped based on their data distributions, and the groups exhibit variations in the number of available clients. We present FLICS-OPT, an algorithm for large-scale federated learning over heterogeneous data distributions, time-varying client availability and further constraints on client participation reflecting, e.g., overall energy efficiency objectives that should be met to achieve sustainable deployment. In particular, FLICS-OPT dynamically learns a selection policy that adapts to client availability patterns and communication constraints, ensuring per-group long-term participation which minimizes the variance inevitably introduced into the learning process by client sampling. We show that for non-convex smooth functions FLICS-OPT coupled with SGD converges at $O(1/\sqrt{T})$ rate, matching the state-of-the-art convergence results which require clients to be always available. We test FLICS-OPT on three realistic federated datasets and show that, in terms of maximum accuracy, FLICS-AVG and FLICS-ADAM outperform traditional FEDAVG by up to 280% and 60%, respectively, while exhibiting robustness in face of heterogeneous data distributions.**

***Index Terms*—Federated learning, intermittency constraints, communication efficiency.**

I. INTRODUCTION

THE aim of cross-device federated learning (FL) is to enable training of a global model $\mathbf{w} \in \mathbb{R}^d$ on generally heterogeneous (i.e., non-i.i.d.) data that is distributed across a potentially vast number of client devices. By securely aggregating local model updates, rather than exchanging potentially massive amounts of possibly personal data, FL systems promote data privacy [1], [2] and offer operational advantages that led to their wide adoption in large-scale system applications by, e.g., Apple [3], Google [4], and Meta [5].

In this paper we consider a large-scale setting where the clients belong to one of M groups according to the distribution of their local data. Specifically, each client in group $j \in [M]$ draws data from distribution $\mathcal{D}_j$, and the objective is to minimize a weighted average loss

$$F(\mathbf{w}) := \sum_{j=1}^{M} p_j F_j(\mathbf{w}), \tag{1}$$

where $F_j(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_j}[f(\mathbf{w}; \xi)]$ denotes the loss of clients in group j, and $\mathbf{p} = (p_1, ..., p_M)$ is a vector on the $(M-1)-$ dimensional simplex representing the weights that the system designer assigns to client groups, e.g., based on the client or data availability across groups or some other design objective. Typically, the objective (1) is minimized using distributed versions of SGD [1], [6] or its more sophisticated variants such as [7], [8].

In practical FL systems, the clients may be intermittent or transient, i.e., and the communication bandwidth may vary with time. These issues impact the performance, especially when the data is heterogeneous and the availability patterns vary across the client groups – in such scenarios, the learned global model is biased towards the groups that enjoy more resources while the users in other groups may find that the performance of the global model on their data is poor. Such a performance imbalance cannot be compensated by tracking individual clients and adjusting their sampling rates because in large systems each client will with high probability be selected by the server at most once.

A strategy that achieves desired long-term per-group client participation while making the most of the communication budget is not obvious; informally, the "best" participation levels avoid biases and minimize model performance variations introduced by client sampling. Moreover, such participation levels are a priori unknown; even if they were known, devising a client sampling policy that would ensure the desired participation under unknown, possibly stochastic, client availability patterns and time-varying communication constraints is challenging.

Efficient client sampling strategies in large-scale federated learning systems provide avenues for responsible energy-efficient gathering of model updates, and thus enable sustainable deployments that allocate resources more fairly and minimize the potential adverse impacts on the environment and society. The basic premise of these strategies is that rather than collecting all data and/or model updates, only those that provide the best value towards expediting the desired learning task should be gathered.

A. Contributions

We propose an algorithm for federated learning with intermittent clients at scale (FLICS-OPT) which employs a client sampling policy that adaptively achieves desirable per-group participation. Specifically, FLICS-OPT facilitates training of

Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA, 78703.

an unbiased model (with respect to weights $\mathbf{p}$) while greedily minimizing σ_T, a measure of model variability over groups after T rounds. We show that in a stationary regime, the achieved long-term per-group participation converges to the values achieving the same σ_T as that achieved by a genie-aided offline policy which assumes full prior knowledge of the availability patterns and communication constraints. We establish the convergence of FL under the proposed sampling policy for non-convex functions and stationary availability processes. We test FLICS-OPT on a synthetic dataset as well as CIFAR100 and EMNIST, and show that it outperforms state-of-the-art sampling schemes across real and simulated availability profiles.

B. Related Work

Significant efforts have been invested in studying client selection policies in FL [9]–[15]. These works use statistical methods to produce unbiased estimates of gradient over subsampled pools of clients [9]–[12], [14], and/or rely on stochastic modeling of the gradient update process [13], [15]. The results of investigating client intermittency in the settings where the client availability is known to follow block-cyclic or i.i.d. patterns were reported in [15]–[17]. Other prior research that explores client selection under intermittency and communication constraints includes [18], which investigates FL in cellular wireless networks, and [19], which presents a selection scheme with availability constraints. However, the framework in [19] is limited to strongly convex functions and does not scale efficiently as it requires the server to track individual clients and maintain counts of their participation in training. These two approaches do not take into account the large-scale regime where a client is likely to participate in training at most once. To our knowledge, FL under time-varying communication constraints and group-specific client availability has not been studied. We also note a complementary line of work on asynchronous methods in FL aiming to address the problems of device heterogeneity and its impact on training speed [20]. There, the client sampling policy is fixed and determined by clients' training speed, and updates are incorporated as they arrive to the server individually [21] or in buffers [22].

Remark 1. Our work is related to clustering [23] and personalization in federated learning including Model-Agnostic Meta-Learning (MAML) [24] and its clustering variants [25], [26], and to fine-tuning over clusters [27]–[29]. This line of research shows that clustering users can be fruitful and feasible in practice. We rely on the similar organization of clients into clusters/groups, but our goal is orthogonal/complementary; in fact, the client-selection strategy that FLICS-OPT learns may enhance the performance of various FL methods (including the aforementioned ones) operating under the real-world constraints imposed by client intermittency and time-varying communication constraints.

II. PRELIMINARIES

Notation. We use $[N]$ to denote the set $[N] = \{1, ..., N\}$, and $g(\cdot)$ to denote an oracle which for the received client's index returns that client's group assignment. We use bold letters to denote vectors. Given $\mathbf{v}(t) \in \mathbb{R}^d$ for $t \geq 0$ and for $0 \leq i \leq j$, $\mathbf{v}(i{:}j)$ denotes the collection of vectors $\mathbf{v}(i), ..., \mathbf{v}(j)$. The notation is summarized in Table II in Section A of Appendix.

A. Federated Learning

Let $\mathcal{X}$ be a data domain and $\mathcal{D}_1, ..., \mathcal{D}_M$ denote M distributions over $\mathcal{X}$. Assume that N clients draw data from one of these M distributions; specifically, let us assume that client $i \in [N]$ draws its data from $\mathcal{D}_{g(i)}$, where $g : [N] \to [M]$ is the mapping that specifies the clients' group assignments. We are interested in minimizing the objective (1) where $\mathbf{p} = (p_1, ..., p_M)$ denote the weights assigned to client groups, $F_j(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_j}[f(\mathbf{w}; \xi)]$ for $j \in [M]$, and f is a loss function.

Remark 2. The above formulation implies that the users which belong to the same group have data generated from the same distribution. The groups have mutually distinct data distributions, which models heterogeneity across the groups; this stands in contrast to the traditional formulation of federated learning where the weights $\mathbf{p}$ are assigned to individual clients, each one with a potentially distinct data distribution. Our framework readily accommodates the latter setting by allowing for single-member groups.

B. System model

We consider a large-scale system characterized by: (i) the participation of a vast number of *intermittent* and/or *transient* client devices that become available for some period of time but eventually leave; (ii) given these dynamics and possibly privacy concerns, the tracking of individual clients is useless and/or undesirable but the tracking of group level availability and participation is feasible; and (iii) the system operates under *soft* constraints (possibly stochastic) on the *average* number of participating clients. To formalize this, let $\mathbf{A}(t) = (A_j(t) : j \in [M])$, where $A_j(t)$ denotes the (possibly) random number of clients available at time t from group j and $K(t)$ denotes a constraint on the average number of clients participating at time t. Given a realization $(\mathbf{a}(t), k(t)) \in \mathcal{C}$ of $(\mathbf{A}(t), K(t))$ where $\mathcal{C}$ is a state space of the process, we consider a *probabilistic sampling* of clients from each group which results in an average number of participating clients $r_j(t)$ from group j; clearly $r_j(t) \leq a_j(t)$. As mentioned, a feasible sampling policy should be such that the overall average number of participating clients satisfies $\sum_{j \in [M]} r_j(t) \leq k(t)$. Definition 1 formalizes the notion of a feasible vector for the average number of participating clients.

Definition 1. *Given $(\mathbf{A}(t), K(t)) = (\boldsymbol{a}, k)$, we say that $\mathbf{r} = [r_1, \ldots, r_M]$ is a* **feasible** *vector for the average number of participating clients at time t if*

$$0 \leq r_j \leq a_j \quad \text{for } j \in [M] \tag{2}$$

$$\sum_{j \in [m]} r_j \leq k. \tag{3}$$

We denote the set of all such feasible vectors $\mathbf{r}$ by $\mathcal{R}(\boldsymbol{a}, k)$.

Definition 2. *A stationary state-dependent probabilistic sampling policy* $\mathbf{f}$ *is such that, for every* $(\boldsymbol{a}, k)$*, it leads to a feasible vector for the average number of participating clients, i.e.,* $\boldsymbol{r}^{\boldsymbol{f}}(\mathbf{a}, k) \in \mathcal{R}(\boldsymbol{a}, k)$.

Definition 3. *If the process* $(\mathbf{A}(t), K(t))_t$ *is stationary with marginal distribution* π *then the* **long-term group participation** *under sampling policy* $\boldsymbol{f}$*, denoted as* $\mathbf{s}^{\boldsymbol{f}}$*, is given by*

$$\mathbf{s}^{\boldsymbol{f}} = \sum_{(\mathbf{a},k)\in\mathcal{C}} \pi(\mathbf{a}, k)\ \boldsymbol{r}^{\boldsymbol{f}}(\mathbf{a}, k). \tag{4}$$

Remark 3. The group participation in Eq. (4) is a counterpart to the *client* participation rate considered in [19]. Utilizing the latter would require learning each client's availability pattern and tracking the client's contributions to the learning process; while this is meaningful in small FL systems where in the absence of resource constraints all clients may participate in all rounds of training, doing so is clearly not feasible when the number of clients is very large and exhibit churn thus some clients may only participate a small number of times in the learning process.

III. The Selection Model

In this section we introduce FLICS-OPT, an algorithm for large-scale federated learning over heterogeneous data distributions and availability patterns, formally stated as Algorithm 1. The intermittently available clients can be organized into groups according to their data distributions; due to the large size of the system, each client likely participates in the training process at most once. The algorithm introduces an online client selection strategy that ensures adherence to the system constraints while providing *per-group* participation minimizing the variance in the learned model introduced by client sampling.

A. Algorithm description

We assume that clients have local access to an oracle $g(\cdot)$ returning their group/cluster assignment (see, e.g., [23], [26] for various methods to achieve this in practice). At a high level, we propose an online strategy at the server which maintains an estimate of the long-term group participation $\hat{s}_k(t)$ for each group of clients, and at each round t chooses a number of clients $r_k(t)$ for each group k that: (i) is feasible for that round (satisfies Eq. (2) and Eq. (3); and (ii) minimizes the variance introduced at that step. Then, the algorithm pings clients in group $j \in [M]$ to respond with probability $r_j(t)/a_j(t)$ and proceeds to update the estimate of the long-term group participation $\widehat{\mathbf{s}}(t)$ using the received number of updates per group. Finally, the server produces an estimate of the gradient for the global model with an importance sampling step that scales samples by the long-term participation estimate $\widehat{\mathbf{s}}(t)$.

We start by describing the proposed client selection policy and the aggregation step, followed by its derivation and the formal statement of the algorithm.

Client selection policy. In large-scale settings, client sampling introduces variance into the stochastic optimization process, impacting convergence rates. This variance depends on group weights p_j in the loss function within our multi-group analysis (Eq. (1)), as demonstrated later. We now outline the client selection policy.. The server initializes $\widehat{\mathbf{s}}(0) = \beta \cdot \mathbf{1} \in \mathbb{R}^M$, where $\beta > 0$. At round t, given an estimate of the long-term participation $\widehat{\mathbf{s}}(t)$, the number of clients selected from each group is such that the variance introduced by client sampling is minimized. To formalize this, below we state $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t))$, an optimization problem that receives as parameters the long-term participation estimate formed in the previous time step, $\widehat{\mathbf{s}}(t-1)$, along with the availabilities and constraints in the current time step, $(\mathbf{a}(t), k(t))$, and returns vector $\mathbf{r}(\mathbf{t})$ used to determine participation probabilities.

$\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t))):$

$$\min_{\mathbf{r}(t)} \sum_{j=1}^{M} \frac{p_j^2}{s_j(t)} \tag{5}$$

$$\text{s.t.} \quad \mathbf{r}(t) \le \mathbf{a}(t), \quad \sum_{j=1}^{M} r_j(t) \le k(t)$$

$$\hat{\mathbf{s}}(t) = \frac{1}{t}\left[(t-1)\widehat{\mathbf{s}}(t-1) + \mathbf{r}(t)\right] \tag{6}$$

It can be shown that the objective function of $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t))$, stated in expression (5), may serve as a proxy for the variance that client sampling introduces in the optimization trajectory (for details, please see Section III-B).

Aggregation. Let us first remind the reader of the fundamental idea behind importance sampling [30], a technique encountered in a variety of fields including federated learning [14], [19], used when generating samples from a desired distribution $\mathbf{p}$ is challenging. Assume that we are interested in estimating the mean of a discrete and finite random variable X under distribution $\mathbf{p} = (p_1, \ldots, p_M)$, but we only have access to samples from a different distribution $\mathbf{s} = (s_1, \ldots, s_M)$. To estimate the mean of X, one can scale the samples and estimate the mean of X as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \frac{p_i}{s_i} y_i$, where y_i, $i \in [n]$ are drawn from $\mathbf{s}$. The same idea may be applied in the context of federated learning: when sampling clients according to the desired per-group participation rates $\mathbf{p}$ is not feasible, one can instead sample clients according to a feasible participation average $\mathbf{s}$ and appropriately scale the client contributions during the aggregation step to ensure no bias.

To summarize, FLICS-OPT starts by observing $(\mathbf{a}(t), k(t))$, the client availability and communication constraints at time t (line 3 in Algorithm 1), and uses them to solve $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$ and determine the average number of sampled clients $\mathbf{r}(t)$ from each group (line 4). The server proceeds by sending $\mathbf{r}(t)$ to clients so that they can locally decide whether to participate or not (line 5), and updating the long-term group participation $\widehat{\mathbf{s}}(t)$ (line 6). The participating clients in $\mathbb{S}_t$ (line 8) receive the initial model $\overline{\mathbf{w}}^t$, and train using a local optimizer (lines 9-13). Finally, the server aggregates the received models, reducing bias through importance sampling

Algorithm 1 Federated learning with intermittent clients at scale (FLICS-OPT).

Input: Parameters: client learning rate η, the number of global rounds T, the number of client local updates T_L, (local) group information assignment $g(\cdot)$.

Output: Global model $\overline{\mathbf{w}}^T$

1: initialize $\overline{\mathbf{w}}_0 \in \mathbb{R}^p$ arbitrarily initialize $\widehat{\mathbf{s}}(0) = \beta \cdot \mathbf{1} \in \mathbb{R}^M$ for small β
2: **for** $t = 1 \to T$ **do**
3: $(\mathbf{a}(t), k(t)) \quad \leftarrow$ observe constraints at time t.
4: $\mathbf{r}(t) \leftarrow \mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$
5: Any available client ℓ responds according to $B_\ell \sim \text{Bernoulli}(r_{g(\ell)}(t)/a_{g(\ell)}(t))$.
6: $\widehat{r}_j(t) \leftarrow$ number of received updates from group $j \in [M]$.
7: $\widehat{\mathbf{s}}(t) = \widehat{\mathbf{s}}(t-1) + \frac{1}{t-1}(\widehat{\mathbf{r}}(t) - \widehat{\mathbf{s}}(t-1))$
8: $\mathbb{S}_t \leftarrow \{\ell : \mathbb{1}_{\{B_\ell = 1\}}\}$
9: **for** Clients $\ell \in \mathbb{S}_t$, in parallel **do**
10: Receive $\overline{\mathbf{w}}^t$ and $\widehat{\mathbf{s}}$ from server.
black
11: $\Delta_\ell^{t+1} \leftarrow$ CLIENTOPT$(\overline{\mathbf{w}}^t, T_L \text{ steps}, \eta_L)$
12: Send $\frac{p_{g(\ell)}}{\widehat{s}_{g(\ell)}(t)} \Delta_\ell^{t+1}$ back to server
13: **end for**black
14: $\Delta^{t+1} \leftarrow \frac{1}{|\mathbb{S}_t|} \sum_{\ell \in \mathbb{S}_t} \frac{p_{g(\ell)}}{\widehat{s}_{g(\ell)}(t)} \Delta_\ell^{t+1}$
15: $\overline{\mathbf{w}}^{t+1} \leftarrow$ SERVEROPT$(\overline{\mathbf{w}}^t, \Delta^{t+1}, \eta)$
16: **end for**

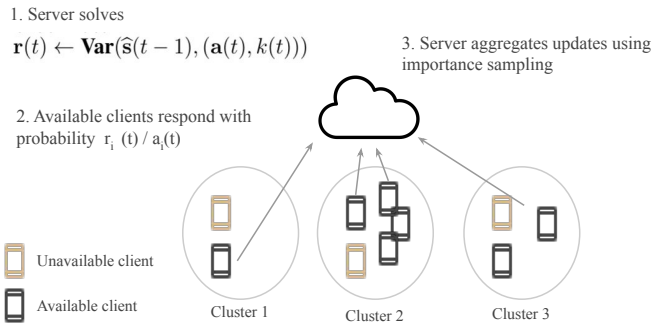

Fig. 1: Illustration of one round of communication of Algorithm 1. (1) Server solves the optimization problem in line 4 and sends the solution to available clients. (2) Available clients in cluster k send an update with probability $\frac{r_k(t)}{a_k(t)}$ (lines 5-12). (3) Server updates the global model with received updates using importance sampling (lines 14-15).

(line 14), and produces a new global model $\overline{\mathbf{w}}^{t+1}$ (line 15) using an optimizer that considers Δ^{t+1} as a pseudo-gradient.

Remark 4. Letting each client decide whether or not to participate (line 6) does not significantly increase computation, memory, nor communication. In turn, it allows individual users to protect their group assignment information from the server, providing additional privacy. Thanks to secure aggregation protocols [31], the server observes only the aggregated results. Alternatively, the server could directly sample a given number of clients from each group at the expense of learning the group assignments of those clients.

Remark 5. *Complexity of* FLICS-OPT. FLICS-OPT builds upon the FedAvg algorithm that already runs in large-scale production environments [30]. There are two additional steps: (i) Line 4 in Algorithm 1, that solves problem $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$; this is a convex optimization problem in M variables that can be efficiently solved at the server side. (ii) Line 5 in Algorithm 1 adds a communication step that requires clients to send one bit to the server with a given (small) probability in order to estimate cluster sizes (see remark 4); this is negligible compared to the gradients that active clients are already communicating.

B. Analysis of the algorithm

We start by stating assumptions required to guarantee convergence of FLICS-AVG, which denotes FLICS-OPT coupled with T_L SGD steps at the clients followed by the averaging at the server, and then proceed to analyze the convergence. We show that the resulting long-term group participation minimizes the variance, and in fact asymptotically provides the best possible variance – the one achieved by the genie-aided policy that has a priori access to the availability patterns and communication constraints rather than seeing them in hindsight.

Algorithm convergence. FLICS-OPT (and its variant being analyzed, FLICS-AVG) operates under arbitrary loss functions and time-varying system availability and communication constraints; the following assumptions are made to facilitate the analysis. Note that Assumption 1 is less restrictive than the assumptions typically made by state-of-the-art FL methods; the latter typically assume (i) all clients are available at all times, (ii) availability is i.i.d. across time and clients, (iii) the communication constraint on the number of selected clients is time-invariant, etc. [6]–[9], [32].

Assumptions 2-4 are regularly made in the federated learning literature [7]–[9], [32]. These assumptions hold in a variety of settings and for a number of common objectives, including logistic regression, generalized linear models, and non-convex $L-$smooth functions.

Assumption 1. *Let $\mathbf{A}(t) = [A_1(t), \dots, A_M(t)]$ be a stochastic process governing per-group availability. Then $(\mathbf{A}(t), K(t))_t$ is an ergodic stochastic process with finite domain $\mathcal{C}$ and stationary distribution $\pi(\mathbf{a}, k)$, $(\mathbf{a}, k) \in \mathcal{C}$.*

Assumption 2. *[$L-$Lipschitz Continuous Gradient.] Let F_i be the local objective functions defined in Eq. (1). For all $\mathbf{v}$ and $\mathbf{w}$, it holds that $\|\nabla F_i(\mathbf{v}) - \nabla F_i(\mathbf{v})\| \leq L\|\mathbf{v} - \mathbf{w}\|_2$, for $i \in [M]$.*

Assumption 3. *[Unbiased stochastic gradient with bounded variance.] If client $i \in [N]$ samples data points ξ independently from distribution $\mathcal{D}_{g(i)}$, then $\mathbb{E}_{\xi\sim\mathcal{D}_{g(i)}}[\nabla f(\mathbf{w}, x)] = \nabla F_{g(i)}(\mathbf{w})$ and $\mathbb{E}_{\xi\sim\mathcal{D}_{g(i)}}[\|\nabla f(\mathbf{w}, \xi) - \nabla F_{g(i)}(\mathbf{w})\|_2^2] \leq \sigma_L^2$.*

Assumption 4. *[The clients' local gradients are bounded.] Let $\mathbf{g}_i^{(t,j)}$ be the gradient of client i, $i \in [N]$, in round t and at local iteration j; then $\|\mathbf{g}_i^{(t,j)}\| \leq G^2$, i.e., the norm of the local gradient is bounded by G.*

Theorem 1 and Corollary 1.1 below establish the convergence rate of FLICS-AVG; the proof of Theorem 1 is in Section B of Appendix. After stating them, we analyze the impact of the proposed client selection policy on the convergence speed.

Theorem 1. *Instate Assumption 1 on the client availabilities and communication constraints, and Assumptions 2-4 on the loss functions (5). Assume the clients locally run T_L steps of SGD and* SERVEROPT$(\overline{\mathbf{w}}^t, \Delta^{t+1}) = \overline{\mathbf{w}}^t + \Delta^{t+1}$. *Let $\mathbf{w}^*$ be the minimizer of the objective (1). Then after running* FLICS-AVG *for T rounds with the initial global model $\mathbf{w}^0$ it holds that*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(\overline{\mathbf{w}}^t)\|^2\right] \leq \frac{f(\overline{\mathbf{w}}^0) - f(\mathbf{w}^*)}{c\eta\eta_L T_L T} + C,$$

where

$$C = \frac{1}{c\hat{\epsilon}}\Big[\frac{L\eta\eta_L\hat{\epsilon}^2\sigma_L^2(\sum_i p_i^2)}{2} + \frac{5L^2\hat{\epsilon}\eta_L^2 T_L M(\sum_i p_i^2))}{2}(\sigma_L^2 + 6T_L\sigma_G^2) + \frac{L\eta\eta_L}{2t_L}(\sigma_L^2 + G^2)\sum_{t=1}^{T}\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}\Big],$$

$\mathbf{s}^* = [s_1^*, \dots, s_M^*]$ is the long-term expected number of the selected clients per group, $\widehat{\mathbf{s}}(t) = [\widehat{s}_1(t), \dots, \widehat{s}_M(t)]$ is defined in line 7 of Algorithm 1 and represents the estimate at time t of $\mathbf{s}^*$, $\mathbf{\Sigma}$ is the correlation matrix of the (random) numbers of selected clients per group, $\hat{\epsilon} = 1 + \epsilon/r_{\min}$ with ϵ such that $|s_k^* - \widehat{s}_k(t)| < \epsilon$ and $r_{\min}$ such that $s_j^* > r_{\min}$ for all $j \in [M]$, and c is a constant.

Corollary 1.1. *Instate the settings of Theorem 1. Letting $\sigma_T(\mathbf{r}(1{:}T)) := \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}$, $\eta_L = \frac{1}{\sqrt{T}T_L L}$, and $\eta = \sqrt{T_L M}$, it holds that*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(\overline{\mathbf{w}}^t)\|^2\right] = O\left(\frac{1 + \sigma_T(\mathbf{r}(1{:}T))}{\sqrt{T}} + \frac{1}{T}\right).$$

Discussion of Corollary 1.1. Corollary 1.1 states that FLICS-AVG converges at a rate bounded by $O(1/\sqrt{T})$, matching the rate guarantees for state-of-the-art federated learning algorithms that make much stronger assumptions on clients availability and system configurations, e.g., they assume all clients are available at all times [32], the ability to track participation of individual clients [19], etc.

We now motivate the objective function of $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$. Let us take a closer look into $\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}$, a term in the expression for $\sigma_T(\mathbf{r}(1{:}T))$ for a fixed t. Assuming that clients from group $i \in [M]$ participate independently at random from each other and across time according to a Bernoulli random variable with parameter r_i (as in Lemma 3.4 in [19]), then the correlation across clients (and, consequently, the groups) is 0, i.e., $\Sigma_{i,j} = 0$ for $i \neq j$. Moreover, $\Sigma_{i,i} = nr_i(1 - r_i)$, and the expected number of selected clients, s_i, is given by $s_i = nr_i$. In this case,

$$\sum_{i,j}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij} = \sum_{i}^{M}\frac{p_i^2}{\hat{s}_i(t)^2}nr_i(1 - r_i) = \sum_{i}^{M}\frac{p_i^2}{\hat{s}_i(t)^2}\left(s_i - \frac{s_i^2}{n}\right) \tag{7}$$

Since $\hat{s}_i(t)$ is the empirical mean, due to the Hoeffding bound it holds with probability $1 - \beta$ that

$$1 - \epsilon \leq \frac{s_i}{\hat{s}_i(t)} \leq 1 + \epsilon$$

for $t = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$. Using this result in Eq. (7), we obtain that with high probability

$$(1 - \epsilon)\sum_{i}^{M}\frac{p_i^2}{\hat{s}_i(t)} - \frac{p_i^2}{n} \leq \sum_{i}^{M}\frac{p_i^2}{\hat{s}_i(t)^2}\left(s_i - \frac{s_i^2}{n}\right) \leq (1 + \epsilon)\sum_{i}^{M}\frac{p_i^2}{\hat{s}_i(t)} - \frac{p_i^2}{n}.$$

Therefore, in this setting, solving $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$ minimizes the variance at each iteration (ignoring constant terms in s_i). The experiments demonstrate efficacy of using $\mathbf{r}(t)$ obtained by solving $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$ in much less restrictive settings than the one described above. Next, we show that this greedy approach, which optimizes the selection rate at each iteration, asymptotically converges to the offline strategy that in hindsight selects a schedule for all T iterations.

Asymptotic optimality of the selection policy. Assume, for the sake of an argument, that we are a priori given the entire realization $(\mathbf{a}(t), k(t))_{t=1}^T$ of the availability and communication constraint processes; let $\mathbf{r}(1{:}T)$ be a sequence of feasible rates for $(\mathbf{a}(t), k(t))_{t=1}^T$. Then, we could identify the full sampling policy $\mathbf{r}(1{:}T)$ by minimizing $\sigma_T(\mathbf{r}(1{:}T))$ rather than iteratively deciding on $\mathbf{r}(t)$ as we do when solving $\mathbf{Var}(\widehat{\mathbf{s}}(t-1), (\mathbf{a}(t), k(t)))$. Furthermore, such prior information would allow us to perform importance sampling using the exact long-term group participation $\mathbf{s}$ instead of the estimates $\widehat{\mathbf{s}}(t)$. Let us formally define such a genie-aided *offline* client sampling policy as the solution to the optimization problem $\mathbf{OffVar}\left((\mathbf{a}(t), k(t))_{t=1}^T\right)$ stated below. Note the distinction between the objective (5) that utilizes importance sampling with different group-participations at each time, and the objective

(8) that uses fixed importance sampling group-participation $\mathbf{s}$ when seeking $\mathbf{r}(1:T)$.

OffVar $\left((\mathbf{a}(t), k(t))_{t=1}^{T}\right)$:

$$\min_{\mathbf{r}(1:T)} \quad \sigma_T^{\text{Off}}(\mathbf{s}(\mathbf{r}(1:T))) := \sum_{i=1}^{M} \frac{p_i^2}{s_i} \tag{8}$$

$$\text{s.t.} \qquad \mathbf{s}(\mathbf{r}(1:T)) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}(t), \quad 0 < r_{\min} \leq \mathbf{s}$$

$$0 \leq \mathbf{r}(t) \leq \mathbf{a}(t), \quad \sum_{k=1}^{M} r_k(t) \leq k(t), \quad t \in [T].$$

Of course, the server can access $(\mathbf{a}(t), k(t))_{t=1}^{T}$ only in hindsight and thus **OffVar** $\left((\mathbf{a}(t), k(t))_{t=1}^{T}\right)$ is not practically feasible. Interestingly, FLICS-AVG is *asymptotically optimal* in the following sense: if $\mathbf{r}^{\text{F}}(1{:}T)$ denote the sampling rates determined by FLICS-AVG, then as T grows, the value of $\sigma_T(\mathbf{r}^{\text{F}}(1{:}T))$ converges to the value achieved by the sampling rates found by the (genie-aided) optimization **OffVar** $\left((\mathbf{a}(t), k(t))_{t=1}^{T}\right)$. This result is formalized in Theorem 2 below; the proof of Theorem 2 is in Sec B of Appendix.

Theorem 2. *Let $(\mathbf{a}(t), k(t))_{t=1}^{T}$ be a realization of the client availabilities and communication constraints meeting Assumption 1. Let $\mathbf{r}^*(1{:}T)$ denote the solution to **OffVar** $\left((\mathbf{a}(t), k(t))_{t=1}^{T}\right)$, and let $\mathbf{r}^{\text{F}}(1{:}T)$ denote the rates determined by running* FLICS-AVG *for T rounds under the assumptions of Corollary 1.1. Then,*

$$\lim_{T\to\infty} \left(\sigma_T(\mathbf{r}^{\text{F}}(1{:}T)) - \sigma_T^{Off}(\mathbf{r}^*(1{:}T))\right) = 0.$$

Theorem 2 states that as T grows, performance of the FLICS-AVG sampling policy approaches the performance of the optimal policy found by an offline algorithm using historic information.

IV. NUMERICAL RESULTS

The numerical results presented in this section demonstrate major performance improvements FLICS-OPT offers over methods that ignore client intermittency/transiency and variable communication constraints. Specifically, we show that FLICS-OPT: (1) achieves better accuracy on a number of datasets for both real and synthetic client availability models; (2) provides a more fair performance than state-of-the-art methods in terms of the worst-case accuracy, i.e., achieves the highest worst accuracy across client groups, (3) converges faster, achieving the maximum accuracy sooner than state-of-the-art methods.

In the experiments below we denote by FLICS-AVG and FLICS-ADAM the counterparts of FLICS-OPT where the server optimizer (Line 15 in Algorithm 1) is replaced with gradient descent and Adam respectively.

A. Datasets and models.

We run experiments on three datasets with varied group structures. First, a clustered version of Synthetic(.5,.5) for logistic regression introduced in [33], with 10 groups consisting of 1000 clients each. For each group $j \in [10]$ we draw $W_j \sim N(0, 0.5)$, $B_j \sim N(0, 0.5)$, and $\mu_j \sim N(B_j, \mathrm{I}_{60})$; the samples for client i in group j, $(X_i^{(j)}, y_i^{(j)})$, are generated as $X_i^{(j)} \sim N(\mu_j, \mathbf{C}^{(j)})$, $y_i^{(j)} = \text{softmax}(w_j^T x_i) + b_j$, $\mathbf{C}^{(j)} = j^{-1.2}\mathrm{I}_{60}$, where I_{60} denotes the 60×60 identity matrix.

Second, we train ResNet-18 for an image recognition task on the federated version of CIFAR100 introduced in [7], where we replace batch normalization by group normalization as in [7]. Specifically, each client draws samples from CIFAR100 according to a sparse multinomial distribution over 20 coarse labels formed using the original 100 classes; the group assignment of each client is specified by the coarse label in the client's dataset with the largest probability of being selected.

Third, we consider an image recognition task with a convolutional neural network on a reprocessed version of EMNIST where we randomly split clients in 10 groups and enforce that the clients in group k are assigned samples with labels k and $\text{mod}(k+1, 10)$. We provide specific network architectures in the appendix.

B. Availability models.

Below are the stochastic processes used to model client availability patterns; for process parameters, please see the appendix. Recall that $A_k(t)$ is the (random) number of available users in group k at time t. We use synthetic and real availability data in our benchmarking experiments. We first describe how the real availability data was collected and then introduce three simulated and two real profiles modeling client availability.

a) Real availability data.: We extracted the real availability data from the dataset introduced by [34], collecting mobile application records for 5,342 users during 2014. Each record provides information for one interaction of a user with an app, including the user identifier, the app used, beginning and ending timestamps, the duration of the interaction (inferred from the timestamps), and the amount of transmitted and received bytes.

We create the availability patterns in the following way. First, we discretize the timestamps at one hour intervals. At hour t, we treat a client as available if: (1) the client is using an app (any app), indicating the client has good network connection, and (2) the client is not heavily using the phone, where the data download and transmission rate under 10MB during a given interval is considered "light". We then calculate each user's average availability by counting the number of slots in which they are available, divided by the overall time they are active. Finally, we can assign to each client in our experiments an availability profile described below as *Real-random* and *Real-correlated* processes.

b) Availability processes:

1) Uniform: $A_k(t)$ is a uniform random variable over the interval $[\alpha_k^{\min}, \alpha_k^{\max}]$.
2) Poisson: $A_k(t)$ is a Poisson random variable with parameter λ_k.
3) Cyclic: $A_k(t)$ is a bi-modal random variable emulating realistic cyclic patters [35]; $A_k(t) \sim \text{Poisson}(\lambda_k^{Day})$ if t is even, and $A_k(t) \sim \text{Poisson}(\lambda_k^{Night})$, otherwise.
4) Real-random: each client receives (uniformly at random) a real availability pattern from the dataset introduced in [34]; for more details, please see the appendix.

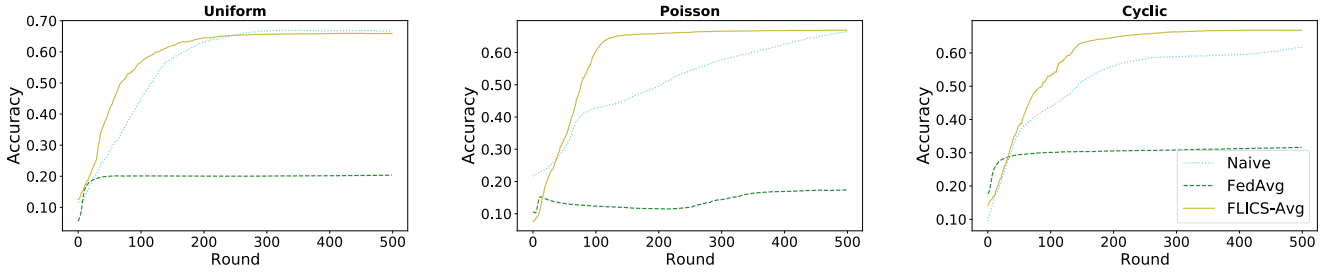

Fig. 2: The convergence under different availability models. FLICS-OPT converges to the optimal value at a faster rate than baselines.

5) Real-correlated: Similar to real-random but we first sort clients according to their availability, and split the resulting list into three equal parts. Then equal number of groups is assigned one of these availability patterns, thus correlating the availability patterns with data distributions.

C. Baselines

Let $k(t)$ be the number of clients selected for training in round t. We use the following baselines:

1) FEDAVG : Server selects $k(t)$ clients from the set of available clients.
2) NAIVE : Server selects at random $\min\{k(t)\cdot p_j, a_j(t)\}$ clients from group $j \in [M]$. This strategy is the optimal one if $p_j k(t) \leq a_j(t)$ for $j \in [M]$.
3) FEDADAM : Server selects $k(t)$ clients from the set of available clients and uses Adam optimizer at the server.

Note that FEDADAM achieves state-of-the-art results on the considered datasets.

D. Results

We ran our experiments on AMD Vega 20 (ROCm) cards.

Accuracy. Table I shows the maximum accuracy achieved by each method over 500 rounds of federated training on Synthetic and EMNIST datasets, and up to 10,000 rounds of federated training on CIFAR100. The methods that sample clients by relying on some information about the relation between group data distributions and client availability, NAIVE and FLICS-OPT, outperform their agnostic counterparts, FEDAVG and FEDADAM. In the uniform setting, NAIVE achieves better performance than FLICS-OPT since the sampling according to p_j is always feasible and is thus the optimal strategy. However, NAIVE ignores instances where the intermittency constraint $r_j(t) \leq a_j(t)$ is active (Poisson and Cyclic models) and selects fewer users from group j than necessary, thus falling behind FLICS-OPT in terms of performance as the latter compensates for the undersampled groups in future rounds. FEDADAM enhances performance of FEDAVG thanks to the momentum variable which helps stabilize the training, while its adaptive learning rate promotes exploration of the optimization space; overall, FLICS-ADAM achieves the best performance.

In addition to the simulated availability models, the tests on Synthetic and EMNIST datasets are also conducted under the real availability models (real-random and real-correlated). Given the relatively short length of the availability pattern time-series from real data ($\sim$1000), the results on CIFAR100 are inconclusive (i.e., CIFAR100 requires longer training). As seen in Table I, FLICS-OPT achieves the best performance in all setting except on the synthetic dataset under real-correlated availability model. In this setting, FEDADAM and FLICS-OPT converge in very few rounds and oscillate around a similar value (further details in the appendix).

Convergence. The plots showing convergence of different schemes on the synthetic data are shown in Fig. 2. In the uniform setting, where it is always feasible to sample proportionally to $\mathbf{p}$, the NAIVE strategy achieves the best performance; however, FLICS-AVG, which has to learn the availability pattern, achieves a comparable accuracy at a faster rate. In all other settings, FLICS-AVG performs the best while NAIVE either converges to a sub-optimal value (cyclic availability model) because it undersamples some groups whenever sampling according to $\mathbf{p}$ is not feasible or achieves the same optimum (Poisson availability model) but much more slowly. FEDAVG always under-performs due to ignoring client intermittency.

Resource constraints. In our experimental setting, all considered methods are given the same communication budget per round. Fig. 3 illustrates how FLICS-OPT helps reduce communication by shortening training time; in particular, FLICS-OPT consistently converges within 200 rounds, accompanied by improved accuracy.

Fairness. FLICS-OPT not only improves the overall maximum accuracy averaged over groups, but also allocates its resources in a more fair way (via more balanced sampling) thus maximizing the minimum accuracy over clusters. We illustrate this effect in Fig. 3 by plotting histograms showing fractions of groups achieving various accuracies on the synthetic data. While FEDAVG leads to a very wide range of group performances, FLICS-AVG and NAIVE succeed in striking a relative balance, with FLICS-AVG clearly outperforming NAIVE in terms of the spread around the mean accuracy. Even for the synthetic dataset under the real-correlated model, where FEDADAM outperforms FLICS-ADAM, FLICS-ADAM achieves a better worst-case performance with a minimum accuracy of 38%, compared to the FEDADAM 's worst group performance of only 30%.

TABLE I: The best accuracy under various simulated availability models after training for over 500 rounds on EMNIST and Synthetic, and 10,000 rounds on CIFAR100.

		Availability model				
		Uniform	Poisson	Cyclic	Real-random	Real-correlated
Synthetic	FedAvg	20.4	17.4	31.6	66.3	59.9
	FLICS-Avg	66.0 (+223.6 %)	66.9 (284.7 %)	66.8 (+111.4 %)	65.9	67.8
	FedAdam	67.0	66.5	66.2	66.6	**66.9**
	FLICS-Adam	**67.6** (0.9%)	**67.4** (1.3 %)	**67.0** (+1.2 %)	**69.4** (+4.2 %)	66.0 (-1.3%)
	Naive	65.5	64.1	52.8	65.6	65.9
EMNIST	FedAvg	64.4	54.8	47.6	89.7	88.2
	FLICS-Avg	93.3 (+44.9 %)	92.7 (+69.3 %)	93.4 (+95.9 %)	93.0 (+3.8 %)	90.9 (+3.0 %)
	FedAdam	70.9	61.8	60.5	93.4	89.0
	FLICS-Adam	**94.4** (+33.1 %)	**94.1** (+52.2 %)	**94.7**(+56.4 %)	**94.4** (+1. %)	**92.7**(+4.1 %)
	Naive	91.7	92.5	93.0	92.8	88.9
CIFAR100	FedAvg	56.2	53.8	53.1	-	-
	FLICS-Avg	57.0 (+1.4 %)	57.1 (+6.1 %)	56.9(+7.1 %)	-	-
	FedAdam	55.7	53.7	53.5	-	-
	FLICS-Adam	**56.9** (+2.1 %)	**56.6** (+5.4 %)	**56.0**(+4.7 %)	-	-
	Naive	55.8	55.0	49.6	-	-

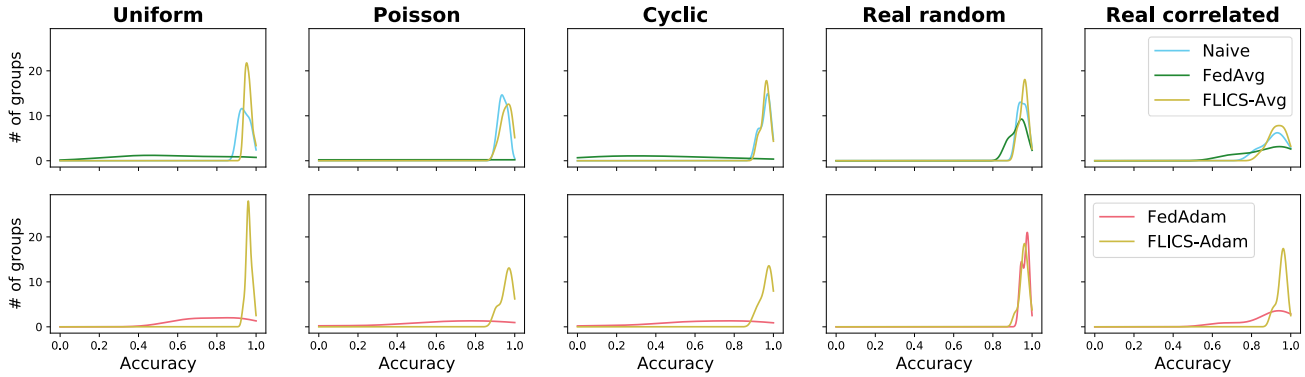

Fig. 3: Histograms showing the fractions of groups achieving various accuracy on EMNIST. FLICS-OPT achieves the highest accuracy in most settings and less variations in performance across different groups.

V. Conclusion

We introduced FLICS-OPT, an algorithm for large-scale federated learning systems where client availability is intermittent and/or exhibits churn/transience. We considered the setting where clients can be partitioned into groups according to their local data distributions; any group's size may change over time, and any client is likely to participate in the learning process no more than once. FLICS-OPT adaptively learns a sampling policy that compensates for variations in availability patterns and communication constraints to achieve asymptotically optimal long-term per-group participation. The combinations of FLICS-OPT with different optimizers, FLICS-AVG and FLICS-ADAM, outperform often significantly their counterparts that ignore client intermittency/transiency and the variations in the availability of communication resources.

FLICS-OPT presents a step towards addressing practical FL design challenges at scale. An interesting avenue of future work includes designing similar strategies to learn personalized models.

Appendix A
Notation

In Table II we introduce the notation used throughout the paper.

Appendix B
Proofs of Theorem 1 and Theorem 2

Theorem (Theorem 1 in main body)**.** *Instate Assumption 1 on the client availabilities and communication constraints, and Assumptions 2-4 on the loss functions (5). Assume the clients locally run T_L steps of SGD and* ServerOpt$(\overline{\mathbf{w}}^t, \Delta^{t+1}) = \overline{\mathbf{w}}^t + \Delta^{t+1}$. *Let $\mathbf{w}^*$ be the minimizer of the objective (1). Then after running* FLICS-AVG *for T rounds with the initial global model $\mathbf{w}^0$ it holds that*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(\overline{\mathbf{w}}^t)\|^2\right] \leq \frac{f(\overline{\mathbf{w}}^0) - f(\mathbf{w}^*)}{c\eta\eta_L T_L T} + C,$$

TABLE II: Frequently used symbols.

Symbol	Definition
$\mathcal{U}$	the set of all clients
$N = \lvert\mathcal{U}\rvert$	the number of clients
M	the number of groups/clusters of clients
$K(t)$	the bound on the number of clients participating in round t
T	total number of rounds
$\mathbf{A}(t)$	the vector of available clients across groups
$\mathbb{S}_t$	set of clients participating at round t
$\pi(\cdot,\cdot)$	stationary distribution of the availability and communication constraint processes
$\mathbf{r}$	the feasible vector for the average number of participating clients
$\mathbf{s}^{\boldsymbol{f}}$	long-term group participation for sampling policy $\boldsymbol{f}$
Δ^{t+1}	pseudo-gradient for server optimizer
Δ_ℓ^{t+1}	client ℓ update at time $t+1$
$\overline{\mathbf{w}}^t$	global model at the beginning of round t
$\mathbf{w}_i^{t+1}$	model at the end of round t at cluster i
$\mathbf{v}_i^{t+1}$	expected update at the end of round t at cluster i
$\overline{\mathbf{v}}^{t+1} = \mathbb{E}_{i\sim\mathcal{P}}\left[\mathbf{v}_i^{t+1}\right] = \sum_{k=1}^N p_i \mathbf{v}_i^{t+1}$	expected global update at the end of round t
$\overline{\mathbf{z}}^{t+1} = \overline{\mathbf{w}}^t + \overline{\mathbf{v}}^{t+1}$	desired global model at the end of round t

where

$$C = \frac{1}{c\hat{\epsilon}}\Big[\frac{L\eta\eta_L\hat{\epsilon}^2\sigma_L^2(\sum_i p_i^2)}{2} + \frac{5L^2\hat{\epsilon}\eta_L^2 T_L M(\sum_i p_i^2))}{2}(\sigma_L^2 + 6T_L\sigma_G^2) + \frac{L\eta\eta_L}{2t_L}\sum_{t=1}^{T}\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}\Big],$$

$\mathbf{s}^* = [s_1^*, \ldots, s_M^*]$ *is the long-term expected number of the selected clients per group,* $\widehat{\mathbf{s}}(t) = [\widehat{s}_1(t), \ldots, \widehat{s}_M(t)]$ *is defined in line 7 of Algorithm 1 and represents the estimate at time* t *of* $\mathbf{s}^*$*,* $\boldsymbol{\Sigma}$ *is the correlation matrix of the (random) numbers of selected clients per group,* $\hat{\epsilon} = 1 + \epsilon/r_{\min}$ *with* ϵ *such that* $|s_k^* - \widehat{s}_k(t)| < \epsilon$ *and* $r_{\min}$ *such that* $s_j^* > r_{\min}$ *for all* $j \in [M]$*, and* c *is a constant.*

Proof of Theorem 1. Let $\overline{\mathbf{w}}^{t+1}$ be the global model after $t+1$ iterations. Let $\mathbb{E}_t[\cdot]$ denote the expectation respect to randomness at round t. By L-smoothness,

$$\begin{aligned}\mathbb{E}_t\left[f(\overline{\mathbf{w}}^{t+1})\right] &\le f(\overline{\mathbf{w}}^t) + \langle\nabla f(\overline{\mathbf{w}}^t), \eta\mathbb{E}_t\left[\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{w}}^t\right]\rangle \\ &\quad + \frac{L}{2}\mathbb{E}_t\left[\|\overline{\mathbf{w}}^{t+1} - \overline{\mathbf{w}}^t\|^2\right] \\ &= f(\overline{\mathbf{w}}^t) + \langle\nabla f(\overline{\mathbf{w}}^t), \eta\mathbb{E}_t\left[\Delta^{t+1}\right]\rangle \\ &\quad + \frac{L\eta^2}{2}\mathbb{E}_t\left[\|\Delta^{t+1}\|^2\right]. \end{aligned} \tag{9}$$

By adding and subtracting $-\hat{\epsilon}\eta\eta_L K\nabla f(\overline{\mathbf{w}}^t)$, where $\hat{\epsilon} > 0$ is a small constant, the last expression on the right-hand side becomes

$$\begin{aligned}&= f(\overline{\mathbf{w}}^t) \\ &\quad + \langle\nabla f(\overline{\mathbf{w}}^t), \eta\mathbb{E}_t\left[\Delta^{t+1}\right] - \hat{\epsilon}\eta\eta_L K\nabla f(\overline{\mathbf{w}}^t) \\ &\quad\quad + \hat{\epsilon}\eta\eta_L K\nabla f(\overline{\mathbf{w}}^t)\rangle + \frac{L\eta^2}{2}\mathbb{E}_t\left[\|\Delta^{t+1}\|^2\right] \\ &= f(\overline{\mathbf{w}}^t) - \hat{\epsilon}\eta\eta_L K\|\nabla f(\overline{\mathbf{w}}^t)\|^2 \\ &\quad + \eta\underbrace{\langle\nabla f(\overline{\mathbf{w}}^t), \mathbb{E}_t\left[\Delta^{t+1}\right] + \hat{\epsilon}\eta_L K\nabla f(\overline{\mathbf{w}}^t)\rangle}_{A_1} \\ &\quad + \frac{L\eta^2}{2}\underbrace{\mathbb{E}_t\left[\|\Delta^{t+1}\|^2\right]}_{A_2}. \end{aligned}$$

Using Lemma 5 and Lemma 7 (stated in Sec C of Appendix) to bound A_1 and A_2, respectively, we obtain

$$\begin{aligned}&\mathbb{E}_t\left[f(\overline{\mathbf{w}}^{t+1})\right] \\ &\le f(\overline{\mathbf{w}}^t) - \eta\eta_L K\hat{\epsilon}(\frac{1}{2} - 15L^2K^2\eta_L^2 M(\sum_i p_i^2))\|\nabla f(\overline{\mathbf{w}}^t)\|^2 \\ &\quad + \frac{\hat{\epsilon}^2\eta^2\eta_L^2 LK\sigma_L^2(\sum_i p_i^2)}{2} \\ &\quad + \frac{5\hat{\epsilon}\eta\eta_L^3 L^2K^2M(\sum_i p_i^2))}{2}(\sigma_L^2 + 6K\sigma_G^2) \\ &\quad - \left(\frac{\hat{\epsilon}^2\eta_L^2\eta^2 L}{2} - \frac{\hat{\epsilon}\eta\eta_L}{2K}\right)\mathbb{E}_t\left[\|\sum_{i=1}^{M} p_i\sum_{k=1}^{K-1}\nabla F_i(\overline{\mathbf{w}}^t)\|^2\right] \\ &\quad + \frac{L\eta^2\eta_L^2}{2}\frac{L\eta^2\eta_L^2}{2}(\sigma_L^2 + G^2)\mathbf{Tr}(\mathbf{Y}_t^T\mathbf{Y}_t\Sigma) \\ &\le f(\overline{\mathbf{w}}^t) - \hat{\epsilon}\eta\eta_L Kc\|\nabla f(\overline{\mathbf{w}}^t)\|^2 + \frac{\eta^2\eta_L^2\hat{\epsilon}^2 LK\sigma_L^2(\sum_i p_i^2)}{2} \\ &\quad + \frac{5\eta\hat{\epsilon}\eta_L^3 L^2K^2M(\sum_i p_i^2))}{2}(\sigma_L^2 + 6K\sigma_G^2) \\ &\quad + \frac{L\eta^2\eta_L^2}{2}(\sigma_L^2 + G^2)\mathbf{Tr}(\mathbf{Y}_t^T\mathbf{Y}_t\Sigma), \end{aligned}$$

where the last inequality holds because $\eta\eta_L < \frac{1}{KL}$ implies $\left(\frac{\hat{\epsilon}^2\eta_L^2\eta^2 L}{2} - \frac{\hat{\epsilon}\eta\eta_L}{2K}\right) > 0$, $\eta_L \leq \frac{1}{\sqrt{30KLM\sum_i p_i^2}}$, and there exist c such that $0 < c < (\frac{1}{2} - 15L^2K^2\eta_L^2 M\sum_i p_i^2))$. Rearranging, taking the sum over $t \in [T-1]$, and taking total expectation,

$$\sum_{t=1}^{T-1} \hat{\epsilon}\eta\eta_L cK\mathbb{E}\left[\|\nabla f(\overline{\mathbf{w}}^t)\|^2\right] \leq f(\overline{\mathbf{w}}^0) - f(\overline{\mathbf{w}}^T) + T\eta\eta_L K\left[\frac{L\eta\eta_L\hat{\epsilon}^2\sigma_L^2(\sum_i p_i^2)}{2} + \frac{5L^2\hat{\epsilon}\eta_L^2 KM(\sum_i p_i^2))}{2}(\sigma_L^2 + 6K\sigma_G^2)\right] + \frac{L\eta^2\eta_L^2}{2}(\sigma_L^2 + G^2)\sum_{t=1}^{T}\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}.$$

Then

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(\overline{\mathbf{w}}^t)\|^2\right] \leq \frac{f(\overline{\mathbf{w}}^0) - f(\mathbf{w}^*)}{c\eta\eta_L KT} + C,$$

where

$$C = \frac{1}{c\hat{\epsilon}}\left[\frac{L\eta\eta_L\hat{\epsilon}^2\sigma_L^2(\sum_i p_i^2)}{2} + \frac{5L^2\hat{\epsilon}\eta_L^2 T_L M(\sum_i p_i^2))}{2}(\sigma_L^2 + 6T_L\sigma_G^2) + \frac{L\eta\eta_L}{2t_L}(\sigma_L^2 + G^2)\sum_{t=1}^{T}\sum_{i,j=1}^{M}\frac{p_i p_j}{\hat{s}_i(t)\hat{s}_j(t)}\Sigma_{ij}\right].$$

Letting $\eta_L = \frac{1}{\sqrt{T}KL}$ and $\eta = \sqrt{KM}$, we obtain the desired convergence of $O\left(\frac{1}{\sqrt{T}} + \frac{1}{T}\right)$.

□

For the completeness of the results discussed in this subsection, we remind the reader of the proposed client selection policy and the off-line client selection policy baseline that Theorem 2 compares. FLICS-AVG determines the average number of sampled clients $\mathbf{r}(t)$ from each group by solving the following optimization problem:

$$\min_{\mathbf{r}(t)} \sum_{j=1}^{M}\frac{p_j^2}{s_j(t)}$$

$$\text{s.t.} \quad \mathbf{r}(t) \leq \mathbf{a}(t), \quad \sum_{j=1}^{M} r_j(t) \leq k(t)$$

$$\mathbf{s}(t) = \frac{1}{t}\left[(t-1)\widehat{\mathbf{s}}(t-1) + \mathbf{r}(t)\right].$$

The genie-aided offline policy instead uses $\mathbf{r}(t)$ obtained by solving the following optimization:

$$\min_{\mathbf{r}(1:T)} \quad \sum_{i=1}^{M}\frac{p_i^2}{s_i}$$

$$\text{s.t.} \quad \mathbf{s}(\mathbf{r}(1:T)) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}(t), \quad 0 < r_{\min} \leq \mathbf{s} \tag{10}$$

$$0 \leq \mathbf{r}(t) \leq \mathbf{a}(t), \quad \sum_{k=1}^{M} r_k(t) \leq k(t), \quad t \in [T].$$

Let $\mathbf{r}^{\mathrm{F}}(1{:}T)$ and $\mathbf{r}^*(1{:}T)$ denote the solutions to the above two optimization problems, respectively. Theorem 2 states that as T grows, performance of the FLICS-AVG sampling policy approaches the performance of the optimal policy found by an offline algorithm using historic information. Recall that we use $\sigma_T^{\mathrm{Off}}(\mathbf{r}^*(1{:}T))$ to denote the optimal value of the objective of the optimization solved by the genie-aided offline policy maker, while $\sigma_T(\mathbf{r}^{\mathrm{F}}(1{:}T))$ still denotes the optimal value of the objective optimized by FLICS-AVG.

Theorem (Theorem 2 in main body)**.** *Let $(\mathbf{a}(t), k(t))_{t=1}^T$ be a realization of the client availabilities and communication constraints meeting Assumption 1. Let $\mathbf{r}^*(1{:}T)$ denote the solution to **OffVar** $\left((\mathbf{a}(t), k(t))_{t=1}^T\right)$, and let $\mathbf{r}^{\mathrm{F}}(1{:}T)$ be the rates determined by running* FLICS-AVG *for T rounds under the assumptions of Corollary 1.1. Then,*

$$\lim_{T\to\infty}(\sigma_T(\mathbf{r}^{\mathrm{F}}(1{:}T)) - \sigma_T^{Off}(\mathbf{r}^*(1{:}T))) = 0.$$

Proof of Theorem 2: We first show that

$$\sigma_T(\mathbf{r}^{\mathrm{F}}(1{:}T)) \to \sigma_T^{\mathrm{Off}}(\mathbf{r}^{\mathrm{F}}(1{:}T)),$$

and then that $\sigma_T^{\mathrm{Off}}(\mathbf{r}^{\mathrm{F}}(1{:}T)) \to \sigma_T^{\mathrm{Off}}(\mathbf{r}^*(1{:}T))$ as $T \to \infty$.

From Lemma 3 it follows that $\lim_{T\to\infty}\widehat{s}_j(T)$ exists and that

$$\lim_{T\to\infty}\widehat{s}_j(T) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_j^{\mathrm{F}}(t). \tag{11}$$

Let $s_j^* = \lim_{T\to\infty}\widehat{s}_j(T)$ and let $\epsilon > 0$. Then there exists T_0 such that

$$\sigma_T(\mathbf{r}^{\mathrm{F}}(1{:}T)) \leq \frac{1}{T}\sum_{t=1}^{T_0}\sum_{i=1}^{M}\frac{p_i}{\widehat{s}_i(t)} + \frac{T-T_0}{T}\sum_{i=1}^{M}\frac{p_i}{s_i^* - \epsilon} = \frac{1}{T}\sigma_{T_0}(\mathbf{r}^{\mathrm{F}}(1:T_0))) + \frac{T-T_0}{T}\sigma_T^{\mathrm{Off}}(\mathbf{s}^* - \epsilon).$$

Given that $\sigma_{T_0}(\mathbf{r}^{\mathrm{F}}(1:T_0)))$ is constant, taking $\epsilon \to 0$ and $T \to \infty$, we obtain the first result,

$$\sigma_T(\mathbf{r}^{\mathrm{F}}(1{:}T)) \to \sigma_T^{\mathrm{Off}}(\mathbf{r}^{\mathrm{F}}(1{:}T)). \tag{12}$$

The analysis that leads to the second claim is inspired by the results on resource allocation in networks [36]. The idea of the proof is to define a penalty function for each cluster, $U_i(r)$, and show through convexity and KKT-conditions that the result given by FLICS-OPT is feasible for **OffVar**.

Due to the optimality of $\mathbf{r}^*(1{:}T)$ we have that

$$\sigma_T^{\mathrm{Off}}(\mathbf{r}^*(1{:}T)) \leq \sigma_T^{\mathrm{Off}}(\mathbf{r}^{\mathrm{F}}(1{:}T)). \tag{13}$$

Note that, by definition, $\mathbf{r}^{\mathrm{F}}(1{:}T)$ is feasible for **OffVar** $\left((\mathbf{a}(t), k(t))_{t=1}^T\right)$. Moreover, the constraints are linear functions and thus convex. Now, note that $U_i^t(r) := \frac{1}{r}$ is a convex function of r, $t \in [T]$, differentiable in an open interval containing $\mathcal{R} := [r_{\min}, r_{\max}]$ for $r_{\min}$ in Eq. (8) and $r_{\max} = \max_t k(t)$ a bound on $K(t)$. Further, being a bounded differentiable function over a bounded domain, $U_i^t(r)$ has a Lipschitz continuous gradient over $\mathcal{R}$.

Combining all of the above, we conclude that **Var**^{t} is a convex optimization problem which satisfies Slater's condition, and thus the following KKT-conditions hold for the optimal

value $\mathbf{r}_t^{\mathrm{F}}$ at iteration t: *There exists non-negative* $\mu_t^* \in \mathbb{R}^{M+1}$ *and* $\gamma_t^* \in \mathbb{R}^M$ *such that for any* $\mathbf{r}$, $t \in [T]$, *and* $i \in [M]$, $j \in [M+1]$,

$$U_i^{t'}(r_i(t)) + \gamma_{i,t}^* - \sum_{j\in[M+1]} \mu_{j,t}^* = 0 \tag{14}$$

$$\mu_{j,t}^*(r_j(t) - a_j(t)) = 0, \quad j \in [M] \tag{15}$$

$$\mu_{M+1,t}^*(\sum_{i=1}^{M} r_j(t) - k(t)) = 0 \tag{16}$$

$$\gamma_{t,i}^* r_i^* = 0. \tag{17}$$

Define Ψ_T as

$$\Psi(\mathbf{r}_{1:T}) := \sum_{i\in[M]} \frac{p_i}{\frac{1}{T}\sum_t r_i(t)} - \sum_t \sum_{j=1}^{M} \frac{\mu_{j,t}^*}{T}(r_j(t) - a_j(t)) + \sum_t \frac{\mu_{j,t}^*}{T}(\sum_{i=1}^{M} r_i(t) - k(t)) \sum_{i,t} \frac{\gamma_{i,t}^*}{T} r_i(t).$$

Ψ_T is the Lagrangian of **OffVar** but evaluated at the Lagrange multipliers defined in Eq. (14)-(17). Given that these sequences of multipliers are non-negative and thanks to the optimality of $\mathbf{r}^*(1{:}T)$, it holds that

$$\sigma_T^{\mathrm{Off}}(\mathbf{r}_{1:T}) \geq \Psi_T(\mathbf{r}_{1:T}). \tag{18}$$

Since Ψ is differentiable and convex,

$$\begin{aligned}
\Psi_T(\mathbf{r}^*(1{:}T)) &\geq \Psi_T(\mathbf{r}^{\mathrm{F}}(1{:}T)) \\
&\quad + \langle \nabla\Psi_T(\mathbf{r}^{\mathrm{F}}(1{:}T)), \mathbf{r}^*(1{:}T) - \mathbf{r}^{\mathrm{F}}(1{:}T)\rangle \\
&= \sum_{i\in[m]} U_i\left(\frac{r_i^{\mathrm{F}}(t)}{T}\right) && (19)\\
&\quad - \sum_{j,t} \frac{\mu_{j,t}^*}{T} c_{t,j}(\mathbf{r}^{\mathrm{F}}(t)) + \sum_{i,t} \frac{\gamma_{i,t}^*}{T} r_i^{\mathrm{F}}(t) && (20)\\
&\quad + \frac{\sum_{t,i}(r_i^*(t) - r_i^{\mathrm{F}}(t))}{T}\left(\frac{U_i'(r_i^{\mathrm{F}}(t))}{T}\right. \\
&\quad \left. - \sum_j \frac{\mu_{t,j}^*[\nabla c_{t,j}(\mathbf{r}^{\mathrm{F}}(1{:}T))]_i}{T} + \frac{\gamma_{t,i}^*}{T}\right). && (21)
\end{aligned}$$

Combining Eq. (18), Eq. (19) and KKT conditions (14)-(17), it follows that

$$\sigma_T^{\mathrm{Off}}(\mathbf{r}^*(1{:}T)) \geq \sum_{i\in[m]} U_i\left(\frac{1}{T}\sum_t r_i^{\mathrm{F}}(t)\right) = \sigma_T^{\mathrm{Off}}(\mathbf{r}^{\mathrm{F}}(1{:}T)).$$

The result follows from combining Eq. (13), Eq. (22), and Eq. (12).

□

Appendix C
Key Lemmas

Lemma 3. *[Theorem 1 in [36]] Let* $(\mathbf{A}(t), K(t))$ *satisfy Assumption 1, let* $\widehat{\mathbf{s}}$ *be defined by Eq. (6), and* $\mathbf{r}^{\mathrm{F}}(1{:}T)$ *be the group participations determined by running* FLICS-AVG *for* T *rounds. Then for all* $i \in [M]$, $\lim_{T\to\infty} \widehat{s}_j(T)$ *exists and*

$$\lim_{T\to\infty} \widehat{s}_j(T) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_j^{\mathrm{F}}(t). \tag{22}$$

Lemma 4. *Let* $\mathbf{s} = \lim_{t\to\infty} \widehat{\mathbf{s}}(t)$ *and* ϵ *be a bound on the error on* $\widehat{\mathbf{s}}$, *i.e.,* $|\hat{s}_i(t) - s_i| \leq \epsilon$. *Instate the setting of Theorem 1. Then the pseudo-gradient* Δ^t *computed at the server (line 14 in Algorithm 1) is nearly unbiased,*

$$\mathbb{E}\left[\Delta^t\right] = (1+\hat{\epsilon})\overline{\mathbf{v}}^{t+1}, \tag{23}$$

where $\hat{\epsilon} = 1 + \frac{\epsilon}{r_{\min}}$ *and* $r_{\min}$ *is defined in Eq. (10).*

Proof. Let $\boldsymbol{f}$ be any density function achieving $\mathbf{s}$ in the statement of the lemma. Expanding the above expression using total expectations over availability states and invoking Definition 3 (i.e., Eq. (4)), we obtain

$$\mathbb{E}_t\left[\Delta^t\right] = \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c)\mathbb{E}\left[\sum_{j\in\mathbb{S}} \frac{p_{g(j)}}{\widehat{\mathbf{s}}(t)_{g(j)}}\Delta_j^t\right]. \tag{24}$$

Let $\mathbf{v}_i^t$ denote the expected update from cluster i at time t (expectation is over the randomness of the availability model π). Since clients' updates are unbiased (due to Assumption 3), $\mathbb{E}\left[\Delta_j^t\right] = \mathbf{v}_{g(j)}^t$, where the expectation is taken over the randomness of minibatches in local training. Since this sampling is independent of the availability model,

$$\mathbb{E}_t\left[\Delta^t\right] = \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c)\mathbb{E}\left[\sum_{j\in\mathbb{S}} \frac{p_{g(j)}}{\widehat{\mathbf{s}}(t)_{g(j)}}\mathbf{v}_{g(j)}^t\right]. \tag{25}$$

Restating equation Eq. (25) in terms of clusters yields $\mathbb{E}_t\left[\Delta^t\right] =$

$$\begin{aligned}
&= \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c)\mathbb{E}\left[\sum_{i=1}^{M}\sum_{j\in\mathbb{S}} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t \mathbb{1}_{\{g(j)=i\}}\right] \\
&= \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c)\sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t \mathbb{E}\left[\sum_{j\in\mathbb{S}} \mathbb{1}_{\{g(j)=i\}}\right].
\end{aligned}$$

By definition, s_i is the expected number of clients participating at round t from cluster i, i.e., $\mathbb{E}\left[\sum_{j\in\mathbb{S}} \mathbb{1}_{\{g(j)=i\}}\right] = s_i$, and thus

$$\mathbb{E}_t\left[\Delta^t\right] = \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c)\sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t s_i.$$

Reorganizing the terms,

$$\begin{aligned}
\mathbb{E}_t\left[\Delta^t\right] &= \sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t s_i \sum_{c\in\mathcal{C}} \pi(C) \int_{c(\mathbf{r})\leq 0} \boldsymbol{f}(\mathbf{r}|c) \\
&= \sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t s_i \cdot 1 = \sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t s_i.
\end{aligned}$$

The last equality follows because we are integrating over the entire density $\boldsymbol{f}$. Finally, by assumption $|\hat{s}_i(t) - s_i| \leq \epsilon$, which implies that $\frac{s_i}{\hat{s}_i(t)} \leq 1 + \frac{\epsilon}{r_{\min}}$, it follows that

$$\sum_{i=1}^{M} \frac{p_i}{\widehat{\mathbf{s}}(t)_i}\mathbf{v}_i^t s_i = (1 + \frac{\epsilon}{r_{\min}})\sum_{i=1}^{M} p_i \mathbf{v}_i^t = (1 + \frac{\epsilon}{r_{\min}})\overline{\mathbf{v}}^t.$$

□

Lemma 5. *Instate the settings of Theorem 1. Then*

$$\begin{aligned}\langle \nabla f(\overline{\mathbf{w}}^t), \quad & \mathbb{E}_t\left[\Delta^{t+1}\right] + \hat{\epsilon}\eta_L K \nabla f(\overline{\mathbf{w}}^t)\rangle \\ &\leq \frac{\hat{\epsilon}\eta_L K}{2}(1 + 30K^2L^2\eta_L^2 m\sum_i p_i^2)\|f(\overline{\mathbf{w}}^t)\|^2 \\ &\quad + \frac{5\hat{\epsilon}\eta_L^3 L^2 K^2 m\sum_i p_i^2}{2}(\sigma_L^2 + 6K\sigma_G^2) \\ &\quad - \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}\left[\|p_i\sum_{j=0}^{K-1}\nabla F_i(\overline{\mathbf{w}})\|^2\right].\end{aligned}$$

Proof. Using the definition of Δ^{t+1} and unbiasedness of local gradients,

$$\begin{aligned}&\langle \nabla f(\overline{\mathbf{w}}^t), \quad \mathbb{E}_t\left[\Delta^{t+1}\right] + \hat{\epsilon}\eta_L K\nabla f(\overline{\mathbf{w}}^t)\rangle \\ &\quad = \big\langle \nabla f(\overline{\mathbf{w}}^t), \\ &\quad \mathbb{E}_t\left[-\eta_L\sum_{i=1}^m\sum_{k=0}^{K-1}\sum_{j\in S}\frac{p_{g(j)}}{\hat{s}_{g(j)}(t)}\nabla F_{g(j)}(\mathbf{w}_j^{(t,k)})\right. \\ &\quad \left. + \ \hat{\epsilon}\eta_L K\sum_{i=1}^M \nabla p_i F_i(\overline{\mathbf{w}}^t)\right]\Big\rangle.\end{aligned}$$

Using Lemma 4, the right-hand side expression above becomes

$$\langle \nabla f(\overline{\mathbf{w}}^t), -\hat{\epsilon}\eta_L\sum_{i=1}^m p_i\mathbb{E}_t\left[\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k}) - \nabla F_i(\overline{\mathbf{w}}^t)\right]\rangle,$$

which after factoring $\sqrt{\hat{\epsilon}\eta_L K}$ to the left term of the dot product readily modifies to

$$\begin{aligned}&\langle\sqrt{\hat{\epsilon}\eta_l K}\nabla f(\overline{\mathbf{w}}^t), \\ &\frac{\sqrt{\hat{\epsilon}\eta_L}}{\sqrt{K}}\sum_{i=1}^m p_i\mathbb{E}_t\left[\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k}) - \nabla F_i(\overline{\mathbf{w}}^t)\right]\rangle.\end{aligned}$$

Using the identity $\langle x, y\rangle = \frac{1}{2}(\|x\|^2 + \|y\|^2 - \|x - y^2\|)$, the latest expression can be written as

$$\begin{aligned}&\frac{\hat{\epsilon}\eta_L K}{2}\|\nabla f(\mathbf{w}^t)\|^2 \\ &\quad + \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}_t\left[\|\sum_{i=1}^m\sum_{k=0}^{K-1}p_i(\nabla F_i(\mathbf{w}_i^{t,k}) - \nabla F_i(\overline{\mathbf{w}}^t))\|^2\right] \\ &\quad - \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}_t\left[\|\sum_{i=1}^m p_1\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k})\|^2\right]. \quad (26)\end{aligned}$$

We proceed by focusing on the middle term in the above expression and applying the identity $\|\sum_{i=1}^n x_i\| \leq n\sum_{i=1}^n\|x_i\|$ as well as the L-Lipschitz assumption to bound (26) as

$$\begin{aligned}&\leq \frac{\hat{\epsilon}\eta_L K}{2}\|\nabla f(\mathbf{w}^t)\|^2 \\ &\quad + \frac{\hat{\epsilon}\eta_L m}{2K}\|\sum_{i=1}^m p_i^2\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F_i(\mathbf{w}_i^{t,k}) - \nabla F_i(\overline{\mathbf{w}}^t))\|^2\right] \\ &\quad - \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}\left[\|\sum_{i=1}^m p_1\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k})\|^2\right] \\ &\leq \frac{\hat{\epsilon}\eta_L K}{2}\|\nabla f(\mathbf{w}^t)\|^2 \\ &\quad + \frac{\hat{\epsilon}\eta_L m L^2}{2K}\|\sum_{i=1}^m p_i^2\sum_{k=0}^{K-1}\mathbb{E}\left[\|\mathbf{w}_i^{t,k} - \overline{\mathbf{w}}^t\|^2\right] \\ &\quad - \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}_t\left[\|\sum_{i=1}^m p_1\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k})\|^2\right] \\ &\leq \hat{\epsilon}\eta_L K\left(\frac{1}{2} + 30K^2L^2\eta_L^2 m\sum_{i=1}^m p_i^2\right)\|\nabla f(\mathbf{w}^t)\|^2 \\ &\quad + \frac{5\hat{\epsilon}\eta_L^3 L^2K^2 m\sum_{i=1}^m p_i^2}{2}(\sigma_L^2 + 6K\sigma_G^2) \\ &\quad - \frac{\hat{\epsilon}\eta_L}{2K}\mathbb{E}\left[\|\sum_{i=1}^m p_i\sum_{k=0}^{K-1}\nabla F_i(\mathbf{w}_i^{t,k})\|^2\right],\end{aligned}$$

where the last inequality follows from Lemma 2 in [32]. □

Lemma 6. *Let* $\mathbf{s}$ *be the long-term user participation produced by* FLICS-OPT*, and* S_i *be the random variable denoting the expected number of users selected from group* i *under any policy with long-term participation* $\mathbf{s}$*. Let* $\hat{\epsilon}$ *be as in Lemma 4 and* $\mathbf{\Sigma}$ *be the covariance matrix of* $S_1, ..., S_M$*. Let* $\mathbf{x}_i$ *be a (potentially random but independent from* S_i*) vector in* $\mathbb{R}^p$ *for* $i \in [M]$*. Then*

$$\begin{aligned}\mathbb{E}\left[\|\sum_{i=1}^m\frac{p_i}{\hat{s}_i(t)}S_i\mathbf{x}_i\|^2\right] &\leq \hat{\epsilon}^2\mathbb{E}\left[\|\sum_i p_i\mathbf{x}_i\|^2\right] \\ &\quad + \sum_{i,j=1}^m\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)}\langle\mathbf{x}_i, \mathbf{x}_j\rangle.\end{aligned}$$

Proof. Recalling that $\mathbb{E}[S_i] = s_i$,

$$\begin{aligned}&\mathbb{E}\left[\|\sum_{i=1}^m\frac{p_i}{\hat{s}_i(t)}S_i\mathbf{x}_i\|^2\right] \\ &= \sum_{i=1}^m\mathbb{E}\left[\frac{p_i^2S_i^2}{\hat{s}_i^2(t)}\|x_i\|^2\right] + \sum_{i\neq j}^m\mathbb{E}\left[\frac{p_ip_jS_iS_j}{\hat{s}_i(t)\hat{s}_j(t)}\langle\mathbf{x}_i, \mathbf{x}_j\rangle\right] \\ &= \sum_{i=1}^m\frac{p_i^2(s_i^2 + \Sigma_{ii})}{s_i^2(t)}\|x_i\|^2 + \sum_{i\neq j}\frac{p_ip_j(s_is_j^2 + \Sigma_{ij})}{\hat{s}_i(t)\hat{s}_j(t)}\langle\mathbf{x}_i, \mathbf{x}_j\rangle \\ &\leq \hat{\epsilon}^2\left(\sum_{i=1}^m p_i^2\|\mathbf{x}_i\|^2 + \sum_{i\neq j}p_ip_j\langle\mathbf{x}_i, \mathbf{x}_j\rangle\right) \\ &\quad + \sum_{i,j=1}^m\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)}\langle\mathbf{x}_i, \mathbf{x}_j\rangle \\ &= \hat{\epsilon}^2\mathbb{E}\left[\|\sum_i p_i\mathbf{x}_i\|^2\right] + \sum_{i,j=1}^M\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)}\langle\mathbf{x}_i, \mathbf{x}_j\rangle.\end{aligned}$$

□

Lemma 7. *Instate the setting of Theorem 1. Then*

$$\mathbb{E}_t\left[\|\Delta^{t+1}\|^2\right] \leq \hat{\epsilon}^2\eta_L^2 K\sigma_L^2\left(\sum_{i=1}^m p_i^2\right) \tag{27}$$

$$+\eta_L^2\hat{\epsilon}^2\mathbb{E}_t\left[\|\sum_{i=1}^m p_i\sum_{k=1}^{K-1}\nabla F(\mathbf{w}_i^{t,k})\|^2\right] \tag{28}$$

$$+\eta_L^2(\sigma_L^2+G^2)\mathbf{Tr}(\mathbf{Y}_t^T\mathbf{Y}_t\mathbf{\Sigma}). \tag{29}$$

Proof. Recall that, as in Lemma 6, S_i denotes the (random) number of users from cluster $i \in [M]$ participating in round t of training. Then

$$\begin{aligned}
\mathbb{E}_t\left[\|\Delta^{t+1}\|^2\right] &= \mathbb{E}_t\left[\|\sum_{j\in\mathbb{S}}\frac{p_{g(j)}}{\hat{s}_{g(j)}(t)}\Delta^t_{g(j)}\|^2\right] \\
&= \mathbb{E}_t\left[\|\sum_{i=1}^M\frac{p_i}{\hat{s}_i(t)}S_i\Delta_i^t\|^2\right] \\
&= \eta_L^2\mathbb{E}_t\left[\|\sum_{i=1}^M\frac{p_iS_i}{\hat{s}_i}\sum_{k=0}^{K-1}g_i^{t,k}\|^2\right] \\
&= \eta_L^2\mathbb{E}_t\left[\|\sum_{i=1}^M\frac{p_iS_i}{\hat{s}_i}\sum_{k=0}^{K-1}g_i^{t,k}-\nabla F(\mathbf{w}_i^{t,k})\|^2\right] \\
&\quad+\hat{\epsilon}\eta_L^2\mathbb{E}_t\left[\|\sum_{i=1}^M\frac{p_iS_i}{\hat{s}_i(t)}\sum_{k=1}^{K-1}\nabla F(\mathbf{w}_i^{t,k})\|^2\right].
\end{aligned}$$

The last expression can be bounded as

$$\begin{aligned}
&\leq \eta_L^2\hat{\epsilon}^2\left(\mathbb{E}_t\left[\|\sum_{i=1}^M p_i\sum_{k=1}^{K-1}g_i^{t,k}-\nabla F(\mathbf{w}_i^{t,k})\|^2\right]\right. \\
&\quad\left.+\mathbb{E}_t\left[\|\sum_{i=1}^M p_i\sum_{k=1}^{K-1}\nabla F(\mathbf{w}_i^{t,k})\|^2\right]\right) \\
&\quad+\eta_L^2\sum_{i,j=1}^M\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)}\langle g_i^{t,k}-\nabla F_i(\mathbf{w}_i^{t,k}), g_j^{t,k}-\nabla F_j(\mathbf{w}_j^{t,k})\rangle \\
&\quad+\eta_L^2\sum_{i,j=1}^M\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)}\langle\nabla F_i(\mathbf{w}_i^{t,k}), \nabla F_j(\mathbf{w}_j^{t,k})\rangle \\
&\leq \hat{\epsilon}^2\eta_L^2K\sigma_L^2\left(\sum_{i=1}^M p_i^2\right)+\eta_L^2\hat{\epsilon}^2\mathbb{E}_t\left[\|\sum_{i=1}^M p_i\sum_{k=1}^{K-1}\nabla F(\mathbf{w}_i^{t,k})\|^2\right] \\
&\quad+\eta_L^2(K\sigma_L^2+G^2)\eta_L^2\sum_{i,j=1}^M\frac{p_ip_j\Sigma_{ij}}{\hat{s}_i(t)\hat{s}_j(t)},
\end{aligned}$$

where we applied Lemma 6 to obtain the first inequality, and the bounded variance (Assumption [3]) and bounded gradient norm (Assumption [4]) assumptions to obtain the second one, thus completing the proof. □

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 2148224 and is supported in part by funds from OUSD R&E, NIST, and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program.

References

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, 2019.

[3] M. Paulik, M. Seigel, H. Mason, D. Telaar, J. Kluivers, R. van Dalen, C. W. Lau, L. Carlson, F. Granqvist, C. Vandevelde, S. Agarwal, J. Freudiger, A. Byde, A. Bhowmick, G. Kapoor, S. Beaumont, Áine Cahill, D. Hughes, O. Javidbakht, F. Dong, R. Rishi, and S. Hung, "Federated evaluation and tuning for on-device personalization: System design applications," 2022. [Online]. Available: https://arxiv.org/pdf/2102.08503.pdf

[4] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[5] B. Stojkovic, J. Woodbridge, Z. Fang, J. Cai, A. Petrov, S. Iyer, D. Huang, P. Yau, A. S. Kumar, H. Jawa *et al.*, "Applied federated learning: Architectural design for robust and efficient learning in privacy aware settings," *arXiv preprint arXiv:2206.00807*, 2022.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 2017, pp. 1273–1282.

[7] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.

[8] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[9] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[10] K. Hsieh, A. Harlap, N. Vijaykumar, D. Konomis, G. R. Ganger, P. B. Gibbons, and O. Mutlu, "Gaia: Geo-distributed machine learning approaching {LAN} speeds," in *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, 2017, pp. 629–647.

[11] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5050–5060.

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

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

[14] E. Rizk, S. Vlaski, and A. H. Sayed, "Optimal importance sampling for federated learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3095–3099.

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

[16] H. Eichner, T. Koren, B. McMahan, N. Srebro, and K. Talwar, "Semi-cyclic stochastic gradient descent," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1764–1773.

[17] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[18] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5136–5151, 2021.

[19] M. Ribero, H. Vikalo, and G. De Veciana, "Federated learning under intermittent client availability and time-varying communication constraints," *arXiv preprint arXiv:2205.06730*, 2022.

[20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[21] Y. Chen, Y. Ning, M. Slawski, and H. Rangwala, "Asynchronous online federated learning for edge devices with non-iid data," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 15–24.

[22] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," *arXiv preprint arXiv:2106.06639*, 2021.

[23] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi, "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3407–3416.

[24] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[25] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 9120–9132.

[26] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, and C. Finn, "Efficiently identifying task groupings for multi-task learning," *arXiv preprint arXiv:2109.04617*, 2021.

[27] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *arXiv preprint arXiv:2006.04088*, 2020.

[28] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, 2020.

[29] M. Xie, G. Long, T. Shen, T. Zhou, X. Wang, J. Jiang, and C. Zhang, "Multi-center federated learning," *arXiv preprint arXiv:2108.08647*, 2021.

[30] T. Kloek and H. K. Van Dijk, "Bayesian estimates of equation system parameters: an application of integration by monte carlo," *Econometrica: Journal of the Econometric Society*, pp. 1–19, 1978.

[31] K. Bonawitz, F. Salehi, J. Konečnỳ, B. McMahan, and M. Gruteser, "Federated learning with autotuned communication-efficient secure aggregation," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1222–1226.

[32] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," in *International Conference on Learning Representations*, 2020.

[33] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *International conference on machine learning*. PMLR, 2014, pp. 1000–1008.

[34] F. A. Silva, A. C. S. A. Domingues, and T. R. M. B. Silva, "Discovering mobile application usage patterns from a large-scale dataset," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 5, jun 2018. [Online]. Available: https://doi.org/10.1145/3209669

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

[36] V. Joseph, G. de Veciana, and A. Arapostathis, "Resource allocation: Realizing mean-variability-fairness tradeoffs," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 19–33, 2014.

[37] S. Alouf, E. Altman, and P. Nain, "Optimal online estimation of the size of a dynamic multicast group," in *Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, 2002, pp. 1109–1118 vol.2.

Mónica Ribero received the B.Sc. degree in mathematics from the Universidad de los Andes, in Bogotá, Colombia 2015. She received her Ph.D. in Electrical and Computer Engineering from the University of Texas at Austin (2022) working on federated learning under privacy and communication constraints and joined Google Research NY as a Research Scientist. She has held research internship positions at Bell Laboratories (2018), CognitiveScale (2019), and Google Research (2020).

Haris Vikalo received the B.S. degree from the University of Zagreb, Croatia, in 1995, the M.S. degree from Lehigh University in 1997, and the Ph.D. degree from Stanford University in 2003, all in electrical engineering. He held a short-term appointment at Bell Laboratories, Murray Hill, NJ, 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 at the California Institute of Technology. Prof. Vikalo has been with the Department of Electrical and Computer Engineering, the University of Texas at Austin, since September 2007. He is a recipient of the 2009 National Science Foundation Career Award. His research interests include signal processing, machine learning, communications and bioinformatics.

Gustavo de Veciana received his Ph.D. in electrical engineering from the U.C. Berkeley in 1993. He is currently a Professor and Associate Chair of the Department of Electrical and Computer Engineering and recipient of the Cockrell Family Regents Chair in Engineering at U.T. Austin. He served as the Director and Associate Director of the Wireless Networking and Communications Group (WNCG) from 2003-2007. His research focuses on the design, analysis and control networks, information theory and applied probability. Current interests include: measurement, modeling and performance evaluation; wireless and sensor networks; architectures and algorithms to design reliable computing and networked systems. Dr. de Veciana is currently an editor at large for the IEEE/ACM Transactions on Networking. He was the recipient of an NSF CAREER Award 1996, and a co-recipient of 7 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 serves on the board of trustees of IMDEA Networks Madrid.