Empowering Federated Learning with Implicit Gossiping: Mitigating Connection Unreliability Amidst Unknown and Arbitrary Dynamics

Ming Xiang, *Student Member, IEEE*, Stratis Ioannidis, *Member, IEEE,* Edmund Yeh, *Senior Member, IEEE,* Carlee Joe-Wong, *Senior Member, IEEE,* Lili Su, *Member, IEEE*

***Abstract*—Federated learning is a popular distributed learning approach for training a machine learning model without disclosing raw data. It consists of a parameter server and a possibly large collection of clients (e.g., in cross-device federated learning) that may operate in congested and changing environments. In this paper, we study federated learning in the presence of stochastic and dynamic communication failures wherein the uplink between the parameter server and client i is on with *unknown* probability p_i^t in round t. Furthermore, we allow the dynamics of p_i^t to be *arbitrary*.**

We first demonstrate that when the p_i^t's vary across clients, the most widely adopted federated learning algorithm, Federated Average (FedAvg), experiences significant bias. To address this observation, we propose Federated Postponed Broadcast (FedPBC), a simple variant of FedAvg. It differs from FedAvg in that the parameter server postpones broadcasting the global model to the clients with active uplinks till the end of each training round. Despite uplink failures, we show that FedPBC converges to a stationary point of the original non-convex objective. On the technical front, postponing the global model broadcasts enables implicit gossiping among the clients with active links in round t. Despite the time-varying nature of p_i^t, we can bound the perturbation of the global model dynamics using techniques to control gossip-type information mixing errors. Extensive experiments have been conducted on real-world datasets over diversified unreliable uplink patterns to corroborate our analysis.

***Index Terms*—Federated learning, communication failures, gossiping, non-convex optimization, fault-tolerance.**

I. Introduction

FEDERATED learning is a distributed machine learning paradigm wherein a parameter server and a collection of end/edge devices (referred to as *clients*) collaboratively train a machine learning model without requiring clients to disclose their local data [2], [3]. Instead of uploading raw data to the parameter server, the clients work at the front line in processing their local data and periodically report their updates to the parameter server, which then effectively aggregates those updates to obtain a new model. The massive system scale and the client heterogeneity in hardware, software, and environments leads to either active [2], [3] or passive [4]–[6] partial client participation, i.e., in each round, the parameter server receives updates from a subset of clients only.

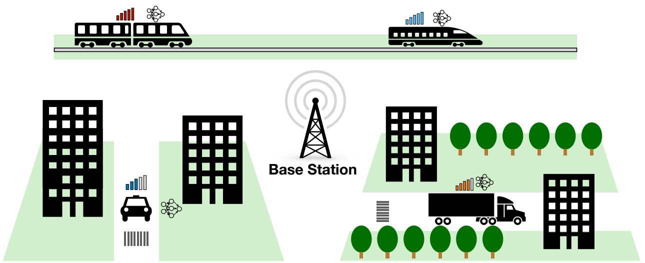


Fig. 1: A federated learning system with moving autonomous vehicles as clients. The signal strength of the vehicles indicates the communication conditions.

Federated learning systems are often deployed in congested and uncontrollable environments with mobile clients such as smartphones and other internet-of-thing devices. Client mobility and environment complexity can result in unreliable communication [3], [7], [8], which may even vary significantly across time and devices. For example, the network connection between a smartphone and a base station may be lost when the smartphone is on a train passing through a tunnel. Popular transportation layer protocols either have an expensive overhead (such as TCP) or are unreliable (such as UDP) [8]. Previous research has demonstrated that unpredictable fluctuations in both the speed and direction of mobile end devices can lead to erratic capacity patterns in 5G links [9]–[11].

Unreliable communication in federated learning systems has not caught attention until recently. Ye et al. [8] assume the communication failures are symmetric with fixed underlying statistics. Time-varying communication constraints are considered in [12], wherein the evolution of the feasible client sets is assumed to follow a homogeneous Markov chain with a steady-state distribution. Yet, as we shall see from the example illustrated in Fig. 1, the assumption of time-invariant communication dynamics easily breaks down when clients are mobile and operate in complex environments. More detailed discussions are reserved in Section II. It is tempting to

A preliminary version of the paper [1] was presented at the IEEE 62nd Conference on Decision and Control 2023, Singapore.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, and Lili Su are with the Department of ECE, Northeastern University, Boston, MA 02115 USA (email: {xiang.mi,l.su}@northeastern.edu; {ioannidis,eyeh}@ece.neu.edu). Carlee Joe-Wong is with the Department of ECE, Carnegie Mellon University, Pittsburgh, PA 15213 USA (email: cjoewong@andrew.cmu.edu).

We gratefully acknowledge the support from the National Science Foundation (NSF) under grants 2106891, 2107062, from the NSF CAREER award under grant 2340482, and from ARO under contract W911NF-23-2-0014.

address dynamic communication capabilities via asynchronous distributed learning, wherein an active client contributes to the global model only when its uplink is on. Unfortunately, to the best of our knowledge, existing literature mostly assumes bounded delay assumption of the uplink availability [13]–[19], which are hard to hold in practical federated learning systems [3], [20]. Often, clients in a federated learning system communicate with the parameter server on their own schedule, which is subject to communication constraints and can have variations due to hardware or software heterogeneity.

In this paper, we study stochastic uplink failures wherein the uplink between the parameter server and client i is active with probability p_i^t in round t. Furthermore, we allow p_i^t to be time-varying and its dynamics to be *unknown* and *arbitrary*. An illustrative example that motivates our problem formulation is shown in Fig. 1. Specifically, fast-moving vehicles quickly pass through a base station's coverage, resulting in frequent handovers. Varying road conditions (e.g., tall buildings, tunnels), traffic densities, and unforeseeable extreme weather can lead to complex dynamics of the connection probabilities. To the best of our knowledge, understanding the convergence of federated learning in the presence of such stochastic uplink failures remains largely under-explored.

Contributions. Our contributions are three-fold:

- We identify simple instances with mild data heterogeneity and show both analytically and numerically that when the p_i^t's are not uniform, *Federated Average* (FedAvg) – the most widely adopted federated learning algorithm – fails to minimize the global objective even for simple convex loss function.
- We propose *Federated Postponed Broadcast* (FedPBC), which differs from FedAvg in that the parameter server postpones broadcasting the global model to the clients with active uplinks till the end of each training round.
  - On the technical front, postponing the global model broadcasts enables implicit gossiping among the clients with active links. Hence, the perturbation caused by non-uniform and time-varying p_i^t can be bounded by leveraging the techniques of controlling information mixing errors.
  - We show in Theorem 1 that, in expectation, FedPBC converges to a stationary point of the non-convex global objective when $p_i^t \geq c$ for an absolute constant c. The staleness of uplink availability is characterized (see Proposition 2). Departing from existing literature, our FedPBC does not require either balanced p_i^t' s, bounded stochastic gradients, or almost surely bounded stochastic gradient noise.
- We validate our analysis empirically on three real-world datasets. Extensive experiments are conducted on both *time-varying* and *time-invariant* Bernoulli, Markovian, and cyclic uplink unreliable patterns.

II. Related Work

In this section, we explore additional related work and present an exhaustive discussion on relevant work mentioned in Section I. The section is divided into two parts: client unavailability and bias correction in distributed learning.

A. Client Unavailability

The communication unreliability addressed in this paper is implicitly linked to client unavailability. The key commonality is that, during failure occurrences, the parameter server cannot receive responses from the involved clients. Prior literature can roughly be categorized into two groups: *known client participation statistics* [2], [4], [21]–[25] and *unknown client participation statistics* [6], [12], [20], [26], [27].

Known client participation statistics. In the seminal works of federated learning [2], [4], the parameter server proactively determines "who to participate" via sampling the clients either uniformly at random or proportionally to clients' local data volume. A more challenging yet practical scenario where the parameter server loses such proactive selection capability is considered in [3]–[5], [28]. To limit the negative impacts of stragglers, the parameter server only waits for a few fastest client responses before moving to the next round. To balance the contribution of active and inactive clients, the parameter server adjusts their aggregation weights according to the corresponding response probabilities, which are assumed to be known. On the other hand, some research aims to *manipulate* client scheduling schemes to either improve communication efficiency or to speed up training, where, at a high level, clients are required to participate whenever the parameter server requests. In contrast, clients are allowed to communicate on their own schedules in our work. To name a few, Perazzone et al. [21] analyze the convergence of FedAvg under time-varying client participation rates. Nevertheless, they assume (1) the participation rates p_i^t's are a known prior and (2) the parameter server controls the participation rates to save communication bandwidth. Chen et al. [24] study a client sampling scheme under which the parameter server only samples the most important updates. Toward this, the parameter server needs to calculate and manipulate the participation rates. Cho et al. [22] devise an adaptive client sampling scheme that non-uniformly selects active clients in each round to accelerate training. Unfortunately, the convergence is up to a non-vanishing error. In another work, Cho et al. [23] study a cyclic participation scheme to accelerate FedAvg training, where the parameter server designs and controls the cyclic participation pattern of the clients. Tang et al. [29] utilize the notion of system-induced bias, where the local data set of active clients does not represent the entire population due to time-varying unbalanced communications. Albeit facing similar time-varying communications, their approach requires, which we do not, the parameter server to select the representative clients strategically.

Unknown client participation statistics. Only a handful of existing works fall under this line of research. Wang and Ji [6] consider structured client unavailability. For the methods in [6] to converge to stationary points, the response rates of the clients need to be "balanced" in the sense that either (1) the p_i^t's are deterministic and satisfy the regularized participation, i.e., there exists $\mu > 0$ such that $\frac{1}{P}\sum_{\tau=1}^{P} p_i^{t_0+\tau} = \mu$ for all clients at all $t_0 \in \{0, P, 2P, \cdots\}$ where P is some carefully chosen integer; or (2) p_i^t's are random and satisfy $\mathbb{E}\left[p_i^t\right] = \mu$ for all clients and sufficiently many rounds. In contrast, we do not require such probabilistic "balanceness". Ribero et al. [12] consider random client availability whose underlying response rates are also heterogeneous and time-varying with unknown dynamics. The key difference from our focus is that the underlying dynamics of p_i^t in [12] is assumed to be Markovian with a unique stationary distribution, which is hard to justify when the dynamics vary significantly. Gu et al. [20] consider general client unavailability patterns for both strongly convex and non-convex global objectives. For non-convex objectives (which is our focus), they require that the consecutive unavailability rounds of a client to be deterministically upper bounded, which does not hold even for the simple uniform and time-invariant response rates. Moreover, they require the noise of the stochastic gradient to be uniformly upper-bounded. Wang and Ji design a lightweight algorithm in a concurrent work [27] to fix FedAvg over non-uniform participation probabilities. However, their algorithm needs a separate online estimation module to adapt clients' aggregation weights to their unavailable durations, while we do not. In addition, they analyze only time-invariant communication probabilities, which are subsumed by our time-varying communication setup.

B. Bias Correction in Distributed Learning

As we will show in Section IV, FedAvg suffers significant bias when the uplinks are non-uniformly available. However, the term bias is not new and has different meanings under different contexts in the field of distributed learning. For example, clients perform multiple local updates to save communication in federated learning before communicating with the parameter server. Yet, bias arises when clients are heterogeneous in the number of local steps [30]. To correct the bias, Wang et al. [30] propose FedNova [30], in which every client participates, and the parameter server normalizes the contribution of different clients by adjusting the aggregation weights according to their numbers of local steps. In fully distributed settings (where no parameter server exists), doubly-stochastic information mixing matrices are critical in ensuring equal contribution among clients. Generally, obtaining doubly-stochastic matrices can be challenging. Push-sum techniques [31], [32] are widely used to address bias that stems from the lack of doubly-stochastic information mixing matrices. However, clients in our problem are only allowed to communicate with the parameter server, rendering direct applications of the techniques impossible. Our setup is orthogonal to them.

III. Problem Formulation

A federated learning system consists of one parameter server and m clients that collaboratively minimize

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F\left(\boldsymbol{x}\right) = \frac{1}{m}\sum_{i\in[m]} F_i\left(\boldsymbol{x}\right), \qquad (1)$$

where $F_i\left(\boldsymbol{x}\right) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\ell_i\left(\boldsymbol{x};\xi_i\right)\right]$ is the local objective, $\mathcal{D}_i$ is the local distribution, ξ_i is a stochastic sample that client i has access to, and ℓ_i is the local loss function. The loss function can be non-convex.

We are interested in solving Eq. (1) over unreliable communication uplinks between the parameter server and the clients. In each round t, the communication uplink between the parameter server and the client i is active with probability p_i^t, which could be simultaneously *time-varying* and is *unknown* to both parameter server and clients. Let $\mathcal{A}^t$ be the set of clients with active uplinks in round t.

Assumption 1 (Threat model)**.** *There exists a $c \in (0,1]$ such that $p_i^t \triangleq \mathbb{E}[\mathbf{1}_{\{i\in\mathcal{A}^t\}}] \geq c$, where the events $i\in\mathcal{A}^t$ are independent across clients $i\in[m]$ and across rounds $t\in[T]$.*

Intuitively, c can be interpreted as one of the system configurations. For our algorithm to work, *neither* the parameter server *nor* clients are required to know c.

Notations. We introduce the additional notations that we will use throughout the paper. For a given vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_2$ defines its l_2 norm. For a given matrix A, $\|A\|_{\mathrm{F}}$ defines its Frobenius norm, and $\lambda_2(A)$ denotes its second largest eigenvalue when A is a square matrix. $\mathbb{R}^d$ defines a d-dimensional vector space. $[m] \triangleq \{1,\cdots,m\}$. $\mathbf{1}_{\{\mathcal{E}\}}$ is an indicator function of event $\mathcal{E}$, i.e., $\mathbf{1}_{\{\mathcal{E}\}} = 1$ when the event $\mathcal{E}$ occurs; $\mathbf{1}_{\{\mathcal{E}\}} = 0$ otherwise. $\mathcal{F}^t$ denotes the sigma-algebra generated by all the randomness up to round t. $O(\cdot)$ is the asymptotic upper bound of a function growth, i.e., $f(n) = O(g(n))$ if there exist constants $c_0 > 0$ and $n_0 \in \mathbb{N}$ such that $f(n) \leq c_0 g(n)$ for all $n \geq n_0$.

IV. A Case Study on the Bias of FedAvg

The heterogeneities in federated learning systems with unreliable uplinks stem from both heterogeneous local data and varying uplink activation probabilities, which together result in a biased objective. In this section, we use a simple quadratic counterexample (a similar setup as in [30]) to illustrate FedAvg fails to minimize the global objective in Eq. (1) when p_i's vary across clients. We numerically observe a similar bias phenomenon when testing other FedAvg-like algorithms such as FedAvg with momentum and FedAvg with two-sided learning rates. Let the local objective $F_i\left(\boldsymbol{x}\right) = \frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{u}_i\right\|_2^2$, where $\boldsymbol{u}_i \in \mathbb{R}^d$ is an arbitrary vector. The corresponding global objective is thus

$$F\left(\boldsymbol{x}\right) = \frac{1}{m}\sum_{i=1}^{m} F_i\left(\boldsymbol{x}\right) = \frac{1}{2m}\sum_{i=1}^{m}\left\|\boldsymbol{x}-\boldsymbol{u}_i\right\|_2^2, \qquad (2)$$

with unique minimizer $\boldsymbol{x}^\star = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{u}_i$.

Proposition 1. *Choose $\boldsymbol{x}^0 = \mathbf{0}$ and $\eta_t = \eta \in (0,1)$ for all t. For a global objective as per Eq. (2) when $p_i^t = p_i$ for all t and under FedAvg with exact local gradients and local computation steps $s \geq 1$, it holds that,*

$$\lim_{T\to\infty}\mathbb{E}\left[\boldsymbol{x}^T\right] = \sum_{i=1}^{m}\frac{p_i\boldsymbol{u}_i\left[1+\sum_{j=2}^{m}(-1)^{j+1}\frac{1}{j}\sum_{S\in\mathcal{B}_j}\prod_{z\in S}p_z\right]}{1-\prod_{i=1}^{m}\left(1-p_i\right)}, \qquad (3)$$

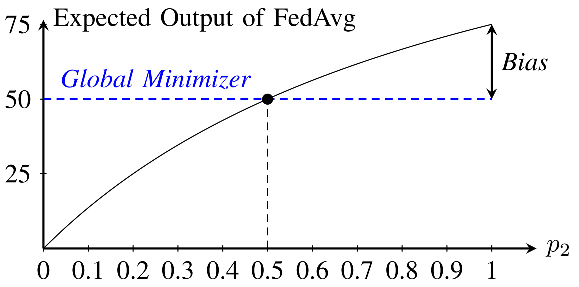


Fig. 2: A visualization of the expected output of FedAvg algorithm with two clients, whose $u_1 = 0, u_2 = 100$ and $p_1 = 0.5$. We vary $p_2 \in [0,1]$ (shown as x-axis). Eq. (3) becomes $\lim_{T\to\infty} \mathbb{E}\left[x^T\right] = (150 \cdot p_2) \,/\, (p_2+1)$. y-axis is the expected output of FedAvg. When $p_2 = 0.5$, FedAvg recovers the global minimizer $(u_1+u_2)/2 = 50$. It can be seen that the expected output of the FedAvg algorithm can deviate far from the global minimizer when $p_1 \neq p_2$.

where $\mathcal{B}_j \triangleq \left\{ S \middle| S \subseteq [m] \setminus \{i\}, |S| = j-1 \right\}$.

It can be checked that if there exist $i, i' \in [m]$ such that $p_i \neq p_{i'}$, then $\lim_{t\to\infty} \mathbb{E}\left[\boldsymbol{x}^t\right] \neq \boldsymbol{x}^*$. In fact, the expected output of FedAvg may be significantly away from $\boldsymbol{x}^\star$ depending on p_i's and $\boldsymbol{u}_i$'s. As illustrated in the scalar example in Fig. 2, overall, the global model in FedAvg deviates away from the global optimum. It is easy to see that the bias only worsens when the connection probabilities p_i's change over time.

On the one hand, when the probability p_i^t's are uniform, (3) reduces to the global optimum $\boldsymbol{x}^\star = \sum_{i=1}^m \boldsymbol{u}_i/m$. In other words, FedAvg recovers the unbiased global optimum when each client's uplink is activated equally often. On the other hand, when clients' local data is i.i.d., e.g., $\boldsymbol{u}_i = \boldsymbol{u}$ for all $i \in [m]$, the expected output of FedAvg recovers the global optimum $\boldsymbol{u}$ under even heterogeneous p_i^t's. This matches the intuition that clients become interchangeable when their local data distributions are homogeneous. We defer the proof to Appendix A.

V. Algorithm: Federated Postponed Broadcast (FedPBC)

In this section, we propose FedPBC (*Federated Postponed Broadcast*, formally described in Algorithm 1) - a simple variant of FedAvg. Recall that $\mathcal{A}^t$ denotes all clients with active communication links in global round t. The stochastic gradient used by client i round t is denoted as $\nabla \ell_i(\boldsymbol{x}_i^{(t,k)}; \xi_i^t)$.

Compared to FedAvg, FedPBC postpones the global model broadcasts to clients in $\mathcal{A}^t$ till the end of each training round. Postponing the global model broadcast introduces some staleness as the clients will start from different $\boldsymbol{x}_i^t$ rather than $\boldsymbol{x}^t$. It turns out that such staleness helps in mitigating the bias caused by non-uniform link activation probabilities. Moreover, the expected staleness is bounded as shown in Proposition 2. Theoretical analysis and numerical results can be found in Sections VI and VII, respectively.

Implicit gossiping among clients in $\mathcal{A}^t$. From line 11 to line 13 of Algorithm 1, via the coordination of the parameter server, the clients in $\mathcal{A}^t$ *implicitly* average their local updates with each other, i.e., there is implicit gossiping among the clients in $\mathcal{A}^t$ at round t. Formally, we are able to construct a mixing matrix $W^{(t)}$ as

$$W_{ij}^{(t)} = \begin{cases} \frac{1}{|\mathcal{A}^t|}, & \text{if } i,j \in \mathcal{A}^t; \\ 1, & \text{if } i=j \text{ and } \{i \notin \mathcal{A}^t\}\,; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The matrix is by definition *doubly-stochastic* and $W^{(t)} = \mathbf{I}$ when $\mathcal{A}^t = \emptyset$ or $|\mathcal{A}^t| = 1$. We further note that this matrix can be *time-varying* since the link activation probabilities p_i^t's can be *time-varying*. As can be seen later, this mixing matrix bridges the gap between local and global model heterogeneity and establishes a consensus among clients. In matrix form, we adopt the following notations.

$$\begin{aligned} \boldsymbol{X}^{(t)} &= \left[\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_m^t\right]; \\ \boldsymbol{G}_0^{(t)} &= \left[s\nabla\ell_1(\boldsymbol{x}_1^{(t,0)}), \cdots, s\nabla\ell_m(\boldsymbol{x}_m^{(t,0)})\right]; \\ \boldsymbol{G}^{(t)} &= \left[\sum_{r=0}^{s-1} \nabla\ell_1(\boldsymbol{x}_1^{(t,r)}), \cdots, \sum_{r=0}^{s-1} \nabla\ell_m(\boldsymbol{x}_m^{(t,r)})\right]; \\ \nabla\boldsymbol{F}^{(t)} &= \left[\nabla F_1(\boldsymbol{x}_1^t), \cdots, \nabla F_m(\boldsymbol{x}_m^t)\right]. \end{aligned}$$

Further, let

$$\bar{\boldsymbol{x}}^t \triangleq \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i^t. \tag{5}$$

Consequently, the consensus error, which measures the distance between the averaged model over all the clients and local

```
Algorithm 1: FedPBC
1  Input: $T$, $x^0$, $s$, $\{\eta_t\}_{t=0,\cdots,T-1}$. The parameter server
   and each client initialize parameter $x^0$;
2  for $t = 0, \cdots, T-1$ do
       /* On the clients.                              */
3      for $i \in [m]$ do
4          $x_i^{(t,0)} = x_i^t$;
5          for $k = 0, \cdots, s-1$ do
6              $x_i^{(t,k+1)} \leftarrow x_i^{(t,k)} - \eta_t \nabla \ell_i(x_i^{(t,k)}; \xi_i^t)$;
7          end
8          $x_i^{t\star} \leftarrow x_i^{(t,s)}$;
9          Report $x_i^{t\star}$ to the parameter server;
10     end
       /* On the parameter server.                     */
11     if $\mathcal{A}^t \neq \emptyset$ then $x^{t+1} \leftarrow \frac{1}{|\mathcal{A}^t|} \sum_{i \in \mathcal{A}^t} x_i^{t\star}$;
12     else $x^{t+1} \leftarrow x^t$ ;
13     for $i \in \mathcal{A}^t$ do $x_i^{t+1} \leftarrow x^{t+1}$ ;
14     else $x_i^{t+1} \leftarrow x_i^t$;
15 end
```

models, can be written in matrix form as (6),

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left\|\bar{\boldsymbol{x}}^t-\boldsymbol{x}_i^t\right\|_2^2 &\triangleq \frac{1}{m}\|\boldsymbol{X}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\\ &=\frac{1}{m}\|\left(\boldsymbol{X}^{(t-1)}-\eta\boldsymbol{G}^{(t-1)}\right)W^{(t-1)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\\ &=\frac{\eta^2}{m}\|\sum_{q=0}^{t-1}\boldsymbol{G}^{(q)}\left(\prod_{l=q}^{t-1}W^{(q)}-\mathbf{J}\right)\|_{\mathrm{F}}^2,\end{aligned} \tag{6}$$

where the last equality follows from the fact that all clients are initiated at the same weights.

VI. Convergence Analysis

A. Assumptions

Before diving into our convergence results, we introduce the regularity assumptions, which are commented towards the end of this subsection.

Assumption 2 (Smoothness)**.** *Each local gradient function $\nabla\ell_i(\theta)$ is L_i-Lipschitz, i.e.,*

$$\left\|\nabla\ell_i(\boldsymbol{x}_1)-\nabla\ell_i(\boldsymbol{x}_2)\right\|_2\le L_i\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\le L\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2,$$

for all $\boldsymbol{x}_1,\boldsymbol{x}_2$, and $i\in[m]$, where $L\triangleq\max_{i\in[m]}L_i$.

Assumption 3 (Bounded Variance)**.** *Stochastic gradients at each client node $i\in[m]$ are unbiased estimates of the true gradient of the local objectives, i.e.,*

$$\mathbb{E}\left[\nabla\ell_i(\boldsymbol{x}_i^t)\mid\mathcal{F}^t\right]=\nabla F_i(\boldsymbol{x}_i^t),$$

and the variance of stochastic gradients at each client node $i\in[m]$ is uniformly bounded, i.e.,

$$\mathbb{E}\left[\left\|\nabla\ell_i(\boldsymbol{x})-\nabla F_i(\boldsymbol{x})\right\|_2^2\mid\mathcal{F}^t\right]\le\sigma^2.$$

Assumption 4. *There exists $F^*\in\mathbb{R}$ such that $F(\boldsymbol{x})\ge F^*$ for all $\boldsymbol{x}\in\mathbb{R}^d$.*

Assumption 5 (Bounded Inter-client Heterogeneity)**.** *We say that local objective function F_i's satisfy (β,ζ)-bounded dissimilarity condition for $\beta,\zeta\ge0$ if*

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x})-\nabla F(\boldsymbol{x})\right\|_2^2\le\beta^2\left\|\nabla F(\boldsymbol{x})\right\|_2^2+\zeta^2. \tag{7}$$

Assumptions, 2, 3 and 4 are standard in federated learning analysis [33]–[35]. Assumption 5 captures the heterogeneity across different users. It is a more relaxed assumption, e.g. than, bounded gradients [22], [26], where they assume $\frac{1}{m}\sum_{i\in[m]}\left\|\nabla F_i(\boldsymbol{x})\right\|_2^2\le\zeta^2$, also than [6], [19], where they assume $\frac{1}{m}\sum_{i\in[m]}\left\|\nabla F_i(\boldsymbol{x})-\nabla F(\boldsymbol{x})\right\|_2^2\le\zeta^2$. When clients have i.i.d. local datasets, it holds for Eq. (7) that $\beta=\zeta=0$ since $F_i=F_j$. Notably, we assume the unbiasedness in Assumption 3 is imposed only at the beginning of each global round.

B. Convergence Results

In this section, we state our key lemmas and our main theorem. All remaining proofs are relegated to Appendix A. Proposition 2 captures the expected staleness of local updates.

Proposition 2. *Define the last active round of the link i as $\tau_i(t)\triangleq\{t'\mid t'<t,i\in\mathcal{A}^{t'}\}$. Given p_i^t such that $p_i^t\ge c$, where c is an absolute constant, we have $\mathbb{E}\left[t-\tau_i(t)\right]\le\frac{1}{c}$.*

Lemma 1 (Lemma 1 in [36])**.** *For $s\ge1$, suppose Assumption 2 holds, we have for all $\boldsymbol{x}\in\mathbb{R}^d$:*

$$\left\|\sum_{k=0}^{s-1}\left[\nabla\ell_i(\boldsymbol{x}^{(t,k)})-\nabla\ell_i(\boldsymbol{x}^t)\right]\right\|_2\le\kappa\eta\binom{s}{2}L_i\left\|\nabla\ell_i(\boldsymbol{x}^t)\right\|_2,$$

where $\kappa\triangleq\max_i\frac{(1+\eta L_i)^s-1-s\eta L_i}{\binom{s}{2}(\eta L_i)^2}$ and monotonically non-decreases with respect to $\eta>0$.

Remark 1. *Lemma 1 comes from a concurrent work [36] and characterizes the perturbation incurred by the multi-step local computation. When $s=1$, i.e., when a client performs only one-step local computation, it holds that $\kappa=0$. For $s\ge2$, we have $\kappa\ge1$. Moreover, due to its monotonicity with respect to η in Lemma 1, κ is bounded from above by an absolute constant when the learning rate η is upper bounded.*

Lemma 2 (Descent Lemma)**.** *Suppose Assumptions 2, 3, and 5 hold. Choose a learning rate η such that $\eta\le\frac{1}{108L^2s^3(\beta^2+1)(1+\kappa^2L^2)}$. When Lipschitz constant $L\ge1$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1})-F(\bar{\boldsymbol{x}}^t)\mid\mathcal{F}^t\right]\le-\frac{\eta s}{3}\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\\ &\quad+\eta s\frac{L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2+\eta^2s^26L\left(\zeta^2+\sigma^2\right)\left(1+\kappa^2L^2\right).\end{aligned}$$

Proof of Lemma 2. By Assumption 2, we have

$$\begin{aligned}&F(\bar{\boldsymbol{x}}^{t+1})-F(\bar{\boldsymbol{x}}^t)\le\left\langle\nabla F(\bar{\boldsymbol{x}}^t),\bar{\boldsymbol{x}}^{t+1}-\bar{\boldsymbol{x}}^t\right\rangle+\frac{L}{2}\left\|\bar{\boldsymbol{x}}^{t+1}-\bar{\boldsymbol{x}}^t\right\|_2^2\\ &=\left\langle\nabla F(\bar{\boldsymbol{x}}^t),-\frac{\eta}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\rangle+\frac{L\eta^2}{2}\left\|\frac{1}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\|_2^2.\end{aligned}$$

Taking expectations with respect to the randomness in the mini-batches at t-th rounds, we have

$$\begin{aligned}&\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1})-F(\bar{\boldsymbol{x}}^t)\mid\mathcal{F}^t\right]\\ &\le\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{x}}^t),-\frac{\eta}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\rangle+\frac{L\eta^2}{2}\left\|\frac{1}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\|_2^2\mid\mathcal{F}^t\right].\end{aligned}$$

For ease of notations, we abbreviate $\nabla\ell_i(\boldsymbol{x}_i^{(t,k)})$ as $\nabla\ell_i^{(t,k)}$.

(a) Bounding $\mathbb{E}[\langle \nabla f(\bar{\boldsymbol{x}}^t), -\frac{\eta}{m}\nabla \boldsymbol{G}^{(t)}\mathbf{1}\rangle \mid \mathcal{F}^t]$.

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{x}}^t), -\frac{\eta}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\rangle \mid \mathcal{F}^t\right] \\
&= -\frac{\eta}{m}\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \sum_{i=1}^{m}\sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\rangle \mid \mathcal{F}^t\right] \\
&= \underbrace{-\frac{s\eta}{m}\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\rangle}_{\text{(A)}} \\
&\quad \underbrace{+\,\mathbb{E}\left[\frac{\eta}{m}\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \sum_{i=1}^{m}s\nabla \ell_i^{(t,0)} - \sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\rangle \mid \mathcal{F}^t\right]}_{\text{(B)}}.
\end{aligned}$$

Term (A) can be bounded as

$$\begin{aligned}
&-s\eta\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\rangle = -\frac{s\eta}{2}\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad + \frac{s\eta}{2}\left\|\nabla F(\bar{\boldsymbol{x}}^t) - \frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\|_2^2 - \frac{s\eta}{2}\left\|\frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\|_2^2 \\
&\le -\frac{s\eta}{2}\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 - \frac{s\eta}{2}\left\|\frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\|_2^2 \\
&\quad + \frac{s\eta L^2}{2m}\sum_{i=1}^{m}\left\|\bar{\boldsymbol{x}}^t - \boldsymbol{x}_i^t\right\|_2^2.
\end{aligned}$$

For term (B), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\eta}{m}\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \sum_{i=1}^{m}s\nabla \ell_i^{(t,0)} - \sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\rangle \mid \mathcal{F}^t\right] \\
&= \frac{\eta}{m}\sum_{i=1}^{m}\left\langle \nabla F(\bar{\boldsymbol{x}}^t), \mathbb{E}\left[s\nabla \ell_i^{(t,0)} - \sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)} \mid \mathcal{F}^t\right]\right\rangle \\
&\overset{\text{(a)}}{\le} \frac{\eta^2 s^2}{2}\left\|\nabla F(\bar{x}^t)\right\|_2^2 \\
&\quad \underbrace{+\,\frac{1}{2ms^2}\sum_{i=1}^{m}\mathbb{E}\left[\left\|s\nabla \ell_i^{(t,0)} - \sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\|_2^2 \Big| \mathcal{F}^t\right]}_{\text{(B.1)}},
\end{aligned}$$

where inequality (a) holds because of Young's inequality. From Lemma 1, we bound term (B.1) as follows

$$\begin{aligned}
&\frac{1}{2ms^2}\sum_{i=1}^{m}\mathbb{E}\left[\left\|s\nabla \ell_i^{(t,0)} - \sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\overset{\text{(b)}}{\le} \frac{1}{2ms^2}\sum_{i=1}^{m}\mathbb{E}\left[\kappa^2\eta^2\binom{s}{2}^2 L^2\left\|\nabla \ell_i^{(t,0)}\right\|_2^2 \mid \mathcal{F}^t\right] \\
&= \frac{\kappa^2\eta^2\binom{s}{2}^2 L^2}{2ms^2}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla \ell_i^{(t,0)} - \nabla F_i(\boldsymbol{x}_i^t) + \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\overset{\text{(c)}}{\le} \kappa^2\eta^2 L^2\sigma^2\frac{s^2}{4} + \frac{\kappa^2\eta^2 s^2 L^2}{4m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \\
&\le \kappa^2\eta^2 s^2 L^2\frac{L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 + \kappa^2\eta^2 s^2 L^2(\zeta^2+\sigma^2) \\
&\quad + \kappa^2\eta^2 s^2 L^2\left(\beta^2+1\right)\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2,
\end{aligned}$$

where inequality (b) follows from Lemma 1, inequality (c) follows from Assumption 3, and the last inequality holds because of Proposition 3. Combing the bounds of terms (A) and (B), we get

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{x}}^t), -\frac{\eta}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\rangle \mid \mathcal{F}^t\right] \\
&\le -\left[\frac{s\eta}{2} - \frac{\eta^2 s^2}{2} - \kappa^2\eta^2 s^2 L^2\left(\beta^2+1\right)\right]\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad - \frac{s\eta}{2}\left\|\frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\|_2^2 + \kappa^2\eta^2 s^2 L^2(\zeta^2+\sigma^2) \\
&\quad + \left(\frac{s\eta L^2}{2m} + \kappa^2\eta^2 s^2 L^2\frac{L^2}{m}\right)\sum_{i=1}^{m}\left\|\bar{\boldsymbol{x}}^t - \boldsymbol{x}_i^t\right\|_2^2. \qquad (8)
\end{aligned}$$

(b) Bounding $\mathbb{E}\left[\left\|\frac{1}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\|_2^2 \mid \mathcal{F}^t\right]$. By adding and subtracting, we get

$$\begin{aligned}
&\left\|\frac{1}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\|_2^2 = \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{s-1}\nabla \ell_i^{(t,k)}\right\|_2^2 \\
&\le 2\underbrace{\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{s-1}\left(\nabla \ell_i^{(t,k)} - \nabla \ell_i^{(t,0)}\right)\right\|_2^2}_{\text{(C)}} + 2\underbrace{\left\|\frac{s}{m}\sum_{i=1}^{m}\nabla \ell_i^{(t,0)}\right\|_2^2}_{\text{(D)}}.
\end{aligned}$$

For term (C), by Lemma 1, we have

$$\begin{aligned}
&\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{s-1}\left(\nabla \ell_i^{(t,k)} - \nabla \ell_i^{(t,0)}\right)\right\|_2^2 \\
&\le \frac{\kappa^2\eta^2 s^4 L^2}{4m}\sum_{i=1}^{m}\|\nabla \ell_i^{(t,0)}\|_2^2 \\
&\le \frac{\kappa^2\eta^2 s^4 L^2}{2m}\left(\sum_{i=1}^{m}\left\|\nabla \ell_i^{(t,0)} - \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 + \sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right) \\
&\overset{\text{(d)}}{\le} \frac{\kappa^2\eta^2 s^4 L^2\sigma^2}{2} + \frac{\kappa^2\eta^2 s^4 L^2}{2m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2,
\end{aligned}$$

where inequality (d) holds because of Assumption 3. For term (D), by Assumption 3, we likewise have

$$\frac{s^2}{m^2}\mathbb{E}\left[\|\sum_{i=1}^{m}\nabla \ell_i^{(t,0)}\|_2^2\Big|\mathcal{F}^t\right] \le \frac{2s^2}{m}\left(\sigma^2 + \sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right).$$

Combing the above upper bounds of (C) and (D) and applying Proposition 3, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{m}\boldsymbol{G}^{(t)}\mathbf{1}\right\|_2^2 \mid \mathcal{F}^t\right] &\le 2s^2\sigma^2\left(\frac{2}{m} + \frac{\kappa^2\eta^2 s^2 L^2}{2}\right) \\
&\quad + 6s^2L^2\left(2 + \frac{\kappa^2\eta^2 s^2 L^2}{2}\right)\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 \\
&\quad + 6s^2\left(\beta^2+1\right)\left(2 + \frac{\kappa^2\eta^2 s^2 L^2}{2}\right)\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad + 6s^2\zeta^2\left(2 + \frac{\kappa^2\eta^2 s^2 L^2}{2}\right). \qquad (9)
\end{aligned}$$

(c) Putting them together. Combining (8) and (9), we get

$$\begin{aligned}
&\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1}) - F(\bar{\boldsymbol{x}}^{t}) \mid \mathcal{F}^t\right] \le \kappa^2\eta^2 s^2 L^2(\zeta^2+\sigma^2) \\
&\quad - \frac{\eta s}{2}\left\|\frac{1}{m}\nabla \boldsymbol{F}^{(t)}\mathbf{1}\right\|_2^2 + \frac{L\eta^2}{2}6s^2\zeta^2\left(2+\frac{\kappa^2L^2}{2}\right) \\
&\quad - \left[\frac{\eta s}{2} - \frac{\eta^2 s^2}{2} - \kappa^2\eta^2 s^2 L^2\left(\beta^2+1\right)\right]\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad + \left(\frac{s\eta L^2}{2m} + \kappa^2\eta^2 s^2\frac{L^4}{m}\right)\sum_{i=1}^m \left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 \\
&\quad + \frac{L\eta^2}{2}6s^2L^2\left(2+\frac{\kappa^2L^2}{2}\right)\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 \\
&\quad + \frac{L\eta^2}{2}6s^2\left(\beta^2+1\right)\left(2+\frac{\kappa^2L^2}{2}\right)\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad + \frac{L\eta^2}{2}2s^2\sigma^2\left(\frac{2}{m}+\frac{\kappa^2L^2}{2}\right).
\end{aligned}$$

Assuming that $\eta \le 1/[108Ls(\beta^2+1)(1+\kappa^2L^2)]$, the above displayed equation can be simplified as

$$\begin{aligned}
&\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1}) - F(\bar{\boldsymbol{x}}^{t}) \mid \mathcal{F}^t\right] \le -\frac{\eta s}{3}\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&\quad + \eta s\frac{L^2}{m}\sum_{i=1}^m \left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 + \eta^2 s^2 6L\left(\zeta^2+\sigma^2\right)\left(1+L^2\kappa^2\right).
\end{aligned}$$

□

The consensus error term $\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2$ in Lemma 2 connects our analysis to the aforementioned W matrix. Let

$$\begin{aligned}
&M^{(t)} \triangleq \mathbb{E}\left[\left(W^{(t)}\right)^2\right], \quad \mathbf{J} \triangleq \frac{1}{m}\mathbf{1}\mathbf{1}^\top; \\
&\rho(t) \triangleq \lambda_2\left(M^{(t)}\right) \quad \text{and } \rho \triangleq \max_t \rho(t).
\end{aligned}$$

Next, we borrow insights from the analysis of gossiping algorithms in the following lemma.

Lemma 3 (Ergodicity)**.** *If $p_i^t \ge c$ for some constant $c \in (0,1)$.*

- *For each $t \ge 1$, it holds that $\rho \le 1 - \frac{c^4[1-(1-c)^m]^2}{8}$;*
- *In the special case of uniform and time-invariant availability, suppose it holds that $|\mathcal{A}^t| = k$ for all $t \ge 0$, the bound can be further tightened as $\rho \le 1 - \frac{c^2}{8}$, where $c \triangleq k/m$.*

(Mixing rate, [37, Lemma 1])**.** *For any matrix $B \in \mathbb{R}^{d\times m}$, it holds that*

$$\mathbb{E}_W\left[\|B\left(\prod_{r=1}^t W^{(r)} - \mathbf{J}\right)\|_{\mathrm{F}}^2\right] \le \rho^t \|B\|_{\mathrm{F}}^2, \tag{10}$$

where $\mathbb{E}_W[\cdot]$ denotes an expectation taken with respect to randomness in $W^{(1)}, \cdots, W^{(t)}$.

Proof of Lemma 3. For ease of exposition, we drop time index t in this proof. We first get the explicit expression for $\mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right]$. Suppose that $\mathcal{A} \neq \emptyset$, we have

$$\begin{aligned}
W_{jj'}^2 &= \sum_{k=1}^m W_{jk}W_{j'k} \\
&= W_{jj}W_{j'j} + W_{jj'}W_{j'j'} + \sum_{k\in[m]\setminus\{j,j'\}} W_{jk}W_{j'k}.
\end{aligned}$$

When $k \neq j$ and $k \neq j'$ by Eq. (4), we have

$$W_{jk}W_{j'k} = \frac{1}{|\mathcal{A}|^2}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}}\mathbf{1}_{\{k\in\mathcal{A}\}}.$$

In addition, we have $W_{jj}W_{j'j} = \frac{1}{|\mathcal{A}|^2}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}}$, and $W_{j'j'}W_{jj'} = \frac{1}{|\mathcal{A}|^2}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}}$. Thus,

- For $j \neq j'$, we have
$$W_{jj'}^2 = \sum_{k=1}^m W_{jk}W_{j'k} = \frac{1}{|\mathcal{A}|}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}};$$
- For $j = j'$, we have
$$W_{jj}^2 = \frac{1}{|\mathcal{A}|}\mathbf{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbf{1}_{\{j\in\mathcal{A}\}}\right).$$

(a) The general case where $p_i^t \ge c$. In the special case where $\mathcal{A} = \emptyset$, we simply have $W = \mathbf{I}$ by the algorithmic clauses. Therefore, $\mathbb{E}\left[W_{jj'} \mid \mathcal{A} = \emptyset\right] \ge 0$ holds for any pair of $j, j' \in [m]$. It follows, by the law of total expectation and for all $j, j' \in [m]$, that

$$\begin{aligned}
\mathbb{E}\left[W_{jj'}\right] &= \mathbb{E}\left[W_{jj'} \mid \mathcal{A} = \emptyset\right]\mathbb{P}\left\{\mathcal{A} = \emptyset\right\} \\
&\quad + \mathbb{E}\left[W_{jj'} \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\} \\
&\ge \mathbb{E}\left[W_{jj'} \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\}.
\end{aligned} \tag{11}$$

- For $j \neq j'$, it holds that
$$\begin{aligned}
\mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right] &= \mathbb{E}\left[\frac{1}{|\mathcal{A}|}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}}\Big|\mathcal{A} \neq \emptyset\right] \\
&\overset{(a)}{\ge} \mathbb{E}\left[\frac{1}{m}\mathbf{1}_{\{j\in\mathcal{A}\}}\mathbf{1}_{\{j'\in\mathcal{A}\}}\Big|\mathcal{A} \neq \emptyset\right] = \frac{p_j p_{j'}}{m} \ge \frac{c^2}{m},
\end{aligned}$$
where (a) holds because $|\mathcal{A}| \le m$;
- For $j = j'$, it holds that
$$\begin{aligned}
&\mathbb{E}\left[W_{jj}^2 \mid \mathcal{A} \neq \emptyset\right] = \mathbb{E}\left[\frac{1}{|\mathcal{A}|}\mathbf{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbf{1}_{\{j\in\mathcal{A}\}}\right)\Big|\mathcal{A} \neq \emptyset\right] \\
&\ge \mathbb{E}\left[\frac{1}{m}\left[\mathbf{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbf{1}_{\{j\in\mathcal{A}\}}\right)\right]\Big|\mathcal{A} \neq \emptyset\right] = \frac{1}{m} \ge \frac{c^2}{m}.
\end{aligned}$$

Recall that $M = \mathbb{E}[W]$. Next, we show that each element of M is lower bounded.

$$M_{jj'} \ge \mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\} \ge \frac{c^2}{m}\left[1 - (1-c)^m\right].$$

We note that $\rho(t) = \lambda_2(M)$, where λ_2 is the second largest eigenvalue of matrix M. A Markov chain with M as the transition matrix is ergodic as the chain is (1) *irreducible*: $M_{jj'} \ge \frac{c^2}{m}\left[1 - (1-c)^m\right] > 0$ for $j, j' \in [m]$ and (2) *aperiodic* (it has self-loops). In addition, W matrix is by definition doubly-stochastic. Hence, M has a uniform stationary distribution $\pi = \frac{1}{m}\mathbf{1}^\top$. Furthermore, the irreducible Markov chain is reversible since it holds for all the states that $\pi_i M_{ij} = \pi_j M_{ji}$. The conductance of a reversible Markov chain [38] with a transition matrix M can be bounded by

$$\begin{aligned}
\Phi(M) &= \min_{\sum_{i\in\mathcal{S}}\pi_i \le \frac{1}{2}} \frac{\pi_i \sum_{i\in\mathcal{S}, j\notin\mathcal{S}} M_{ij}}{\sum_{i\in\mathcal{S}}\pi_i} \\
&\ge \frac{\left(\frac{c}{m}\right)^2\left[1 - (1-c)^m\right]|\mathcal{S}|\,|\bar{\mathcal{S}}|}{\frac{|\mathcal{S}|}{m}} = \frac{c^2\left[1 - (1-c)^m\right]}{m}\left|\bar{\mathcal{S}}\right|,
\end{aligned}$$

where $\left|\bar{\mathcal{S}}\right| = m - |\mathcal{S}| \geq \frac{m}{2}$. From Cheeger's inequality, we know that $\frac{1-\lambda_2}{2} \leq \Phi(M) \leq \sqrt{2\left(1-\lambda_2\right)}$. Finally, we have

$$\Phi(M) \geq \frac{c^2\left[1-(1-c)^m\right]}{m}\left|\bar{\mathcal{S}}\right| \geq \frac{c^2\left[1-(1-c)^m\right]}{2}.$$

Thus, $\rho(t) = \lambda_2 \leq 1 - \frac{\Phi^2(M)}{2} \leq 1 - \frac{c^4\left[1-(1-c)^m\right]^2}{8}$.

(b) Select k clients uniformly at random. In each round, the server picks k clients uniformly at random. Consequently, different from the general case where $|\mathcal{A}|$ is a random variable, it holds that $|\mathcal{A}| = k$ and $\mathcal{A} \neq \emptyset$. In addition, $c \triangleq \frac{k}{m}$. After a similar argument as in the first case, it holds that $M_{jj'} \geq \frac{c^2}{k}$. The conductance of the reversible Markov chain with a transition matrix M can be bounded by $\Phi(M) \geq \frac{c^2}{k}\left|\bar{\mathcal{S}}\right| \geq \frac{c}{2}$. Finally, we have $\rho(t) = \lambda_2 \leq 1 - \frac{\Phi^2(M)}{2} \leq 1 - \frac{c^2}{8}$. □

Inequality (10) from [37, Lemma 1] enables us to bound the consensus error term $\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2$ and says that the spectral norm ρ must be less than 1 to ensure a bounded error, which is crucial for the objectives to reach a stationary point. Fortunately, we show that, under our uplink availability assumption, $\rho < 1$ in Lemma 3.

Lemma 4 (Consensus Error)**.** *Suppose Assumptions 2, 3, and 5 hold. Choose a learning rate η such that $\eta \leq \frac{1-\sqrt{\rho}}{108L^2s^3(\beta^2+1)(1+\kappa^2L^2)}$. When Lipschitz constant $L \geq 1$, it holds that*

$$\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{X}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] \leq \frac{12\rho\sigma^2}{(1-\sqrt{\rho})^2}\eta^2 s^2$$

$$+ \frac{54\rho\zeta^2}{(1-\sqrt{\rho})^2}\eta^2 s^2 + \frac{54(\beta^2+1)\rho\eta^2 s^2}{(1-\sqrt{\rho})^2}\frac{1}{mT}\sum_{t=0}^{T-1}\|\nabla F(\bar{\boldsymbol{x}}^t)\|_{\mathrm{F}}^2.$$

Proof of Lemma 4. Define $\Delta\boldsymbol{G}^{(r)} \triangleq \boldsymbol{G}^{(r)} - \boldsymbol{G}_0^{(r)}$ and $A_{r,t} \triangleq \prod_{\ell=r}^t W^{(\ell)} - \mathbf{J}$. The consensus error can be rewritten as

$$\|\boldsymbol{X}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2 = \|(\boldsymbol{X}^{(t-1)} - \eta\boldsymbol{G}^{(t-1)})W^{(t-1)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2$$

$$= \|-\eta\sum_{q=0}^{t-1}\boldsymbol{G}^{(q)}A_{q,t-1}\|_{\mathrm{F}}^2 \leq 3\eta^2\underbrace{\|\sum_{q=0}^{t-1}\Delta\boldsymbol{G}^{(q)}A_{q,t-1}\|_{\mathrm{F}}^2}_{\text{(A)}}$$

$$+ 3\eta^2\underbrace{\|\sum_{q=0}^{t-1}\left(\boldsymbol{G}_0^{(q)} - s\nabla\boldsymbol{F}^{(q)}\right)A_{q,t-1}\|_{\mathrm{F}}}_{\text{(B)}}$$

$$+ 3\eta^2 s^2\underbrace{\|\sum_{q=0}^{t-1}\nabla\boldsymbol{F}^{(q)}A_{q,t-1}\|_{\mathrm{F}}^2}_{\text{(C)}}, \tag{12}$$

where the second equality follows from the fact that all clients are initiated at the same weights.

(a) Bounding $\mathbb{E}\left[\text{(A)}\right]$. The term (A) in Eq. (12) arises from multiple local steps. We have,

$$\mathbb{E}\left[\text{(A)}\right] \overset{\text{(a)}}{\leq} \sum_{q=0}^{t-1}\rho^{t-q}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$+ \sum_{q=0}^{t-1}\sum_{p=0,p\neq q}^{t-1}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(p)}A_{p,t-1}\|_{\mathrm{F}}\|\Delta\boldsymbol{G}^{(q)}A_{q,t-1}\|_{\mathrm{F}}\right]$$

$$\overset{\text{(b)}}{\leq} \sum_{q=0}^{t-1}\rho^{t-q}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$+ \sum_{q=0}^{t-1}\sum_{p=0,p\neq q}^{t-1}\frac{\sqrt{\rho}^{2t-p-q}}{2}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(p)}\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right],$$

where inequality (a) follows from (10), inequality (b) holds because of Young's inequality. Next, we bound the second term. it follows that

$$\sum_{q=0}^{t-1}\sum_{p=0,p\neq q}^{t-1}\frac{\sqrt{\rho}^{2t-p-q}}{2}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(p)}\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq \sum_{q=0}^{t-1}\sum_{p=0}^{t-1}\frac{\sqrt{\rho}^{2t-p-q}}{2}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(p)}\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq \frac{\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right].$$

In addition, since $\rho < 1$, it holds that $\rho^{t-q} \leq \sqrt{\rho}\rho^{\frac{t-q}{2}}$ for any $q \leq t-1$. Thus, we have

$$\mathbb{E}\left[\text{(A)}\right] \leq \sqrt{\rho}\sum_{q=0}^{t-1}\rho^{\frac{t-q}{2}}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$+ \frac{\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq \frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\left(\boldsymbol{G}^{(q)} - \boldsymbol{G}_0^{(q)}\right)\|_{\mathrm{F}}^2\right]. \tag{13}$$

It remains to bound $\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right]$,

$$\mathbb{E}\left[\|\Delta\boldsymbol{G}^{(q)}\|_{\mathrm{F}}^2\right] \overset{\text{(c)}}{\leq} \kappa^2\eta^2 s^4 L^2\mathbb{E}\left[\|\boldsymbol{G}_0^{(q)} - s\nabla\boldsymbol{F}^{(q)} + s\nabla\boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq 2\kappa^2\eta^2 s^4 L^2\mathbb{E}\left[\|\boldsymbol{G}_0^{(q)} - s\nabla\boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$+ 2\kappa^2 s^2\eta^2 s^4 L^2\mathbb{E}\left[\|\nabla\boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq 2\kappa^2 s^2\eta^2 s^4 L^2 m\sigma^2 + 2\kappa^2 s^2\eta^2 s^4 L^2\mathbb{E}\left[\|\nabla\boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right],$$

where inequality (c) follows from Lemma 1, adding and subtracting. Thus,

$$\mathbb{E}\left[\text{(A)}\right] \leq \frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\boldsymbol{G}^{(q)} - \boldsymbol{G}_0^{(q)}\|_{\mathrm{F}}^2\right]$$

$$\leq \frac{4\kappa^2 s^2\eta^2 s^4 L^2 m\sigma^2\rho}{\left(1-\sqrt{\rho}\right)^2}$$

$$+ \frac{4\kappa^2 s^2\eta^2 s^4 L^2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\nabla\boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right].$$

(b) Bounding $\mathbb{E}\left[(\mathrm{B})\right]$.

$$\mathbb{E}\left[(\mathrm{B})\right] \leq \sum_{q=0}^{t-1} \rho^{t-q} \mathbb{E}\left[\| \left(\boldsymbol{G}_0^{(q)} - s\nabla \boldsymbol{F}^{(q)}\right) \|_{\mathrm{F}}^2\right] \leq \frac{\rho m s^2 \sigma^2}{1-\rho}.$$

(c) Bounding $\mathbb{E}\left[(\mathrm{C})\right]$. Use a similar derivation as in (13), and we get

$$\mathbb{E}\left[(\mathrm{C})\right] \leq \frac{2\sqrt{\rho}}{1-\sqrt{\rho}} \sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} \mathbb{E}\left[\|\nabla \boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right].$$

Furthermore, we have

$$\sum_{t=0}^{T-1}\sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} \mathbb{E}\left[\|\nabla \boldsymbol{F}^{(q)}\|_{\mathrm{F}}^2\right] = \sum_{t=0}^{T-2} \mathbb{E}\left[\|\nabla \boldsymbol{F}^{(t)}\|_{\mathrm{F}}^2\right] \sum_{q=1}^{T-1-t} \sqrt{\rho}^q$$
$$\leq \frac{\sqrt{\rho}}{\left(1-\sqrt{\rho}\right)} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla \boldsymbol{F}^{(t)}\|_{\mathrm{F}}^2\right].$$

(d) Putting them together.

$$\frac{1}{mT}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\boldsymbol{X}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] \leq 3\eta^2 s^2 \sigma^2 \frac{\rho\left(1+\kappa^2\eta^2 s^4 L^2\right)}{\left(1-\sqrt{\rho}\right)^2}$$
$$+ \left(\frac{\kappa^2\eta^2 s^4 L^2}{2} + 1\right) \frac{6\eta^2 s^2 \rho}{mT\left(1-\sqrt{\rho}\right)^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla \boldsymbol{F}^{(t)}\|_{\mathrm{F}}^2\right]$$
$$\overset{(\mathrm{d})}{\leq} \frac{9\rho}{(1-\sqrt{\rho})^2}\eta^2 s^2 \frac{1}{mT}\sum_{t=0}^{T-1} \|\nabla \boldsymbol{F}^{(t)}\|_{\mathrm{F}} + \frac{6\rho\sigma^2}{(1-\sqrt{\rho})^2}\eta^2 s^2,$$

where we assume that $\eta \leq \frac{1}{s^2\kappa L}$ in inequality (d). Choosing $\eta \leq \frac{1-\sqrt{\rho}}{6Ls}$ and by Proposition 3, we have

$$\frac{1}{mT}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\boldsymbol{X}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] \leq \frac{12\rho\sigma^2}{(1-\sqrt{\rho})^2}\eta^2 s^2$$
$$\frac{54(\beta^2+1)\rho\eta^2 s^2}{(1-\sqrt{\rho})^2}\frac{1}{mT}\sum_{t=0}^{T-1} \|\nabla \boldsymbol{F}(\bar{\boldsymbol{x}}^t)\|_{\mathrm{F}}^2 + \frac{54\rho\zeta^2}{(1-\sqrt{\rho})^2}\eta^2 s^2.$$

□

Our proof of Lemma 4 shares a similar sketch as that in [37], yet with nontrivial adaptation to account for multiple local updates and the fact the stochastic gradients at a client within each round are *not independent*. Plugging Lemma 4 into Lemma 2, we obtain the main Theorem 1.

Theorem 1. *Suppose Assumptions 2, 3, 4, and 5 hold. Choose a learning rate* η *such that* $\eta \leq \frac{1-\sqrt{\rho}}{108L^2s^3(\beta^2+1)(1+\kappa^2L^2)}$. *When Lipschitz constant* $L \geq 1$*, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \leq \frac{6\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta s T}$$
$$+ 54\eta s L\left(\kappa^2 L^2 + 1 + \frac{1}{1-\sqrt{\rho}}\right)\left(\sigma^2+\zeta^2\right).$$

Corollary 1. *Suppose Assumptions Assumption 2, 3, 4, and 5 hold. Choose* $\eta = 1/\sqrt{T}$*, where* $T \geq \left(108L^2s^3(\beta^2+1)(1+\kappa^2L^2)/\left(1-\sqrt{\rho}\right)\right)^2$. *When Lipschitz constant* $L \geq 1$*, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \leq \frac{6\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{s\sqrt{T}}$$
$$+ 54\frac{sL}{\sqrt{T}}\left(\kappa^2 L^2 + 1 + \frac{1}{1-\sqrt{\rho}}\right)\left(\sigma^2+\zeta^2\right).$$

Remark 2. *Here, we remark on Theorem 1:*
(1) **On the structures.** *The assumption that Lipschitz constant* $L \geq 1$ *is for simplifying the upper bound of* η *only, which, notably, can be readily relaxed but at a cost of a much more sophisticated learning rate condition. The second term stems from noisy stochastic gradients (Assumption 3) and inter-client gradient heterogeneity (Assumption 5).*
(2) **On stationary points of** F**.** *Theorem 1 says that* $\bar{\boldsymbol{x}}^t$ *in FedPBC converges to a stationary point of* F *(non-convex) at a rate of* $1/\sqrt{T}$*. In sharp contrast, Proposition 1 dictates that the expected output of FedAvg converges to a point that could be far away from the true optimum depending on the interplay between* p_i^t*'s and data heterogeneity.*
(3) **On the role of the probability lower bound** c**.** *A larger* c *results in a smaller* ρ *and thus a tighter bound on* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\bar{\boldsymbol{x}}^t\right)\right\|_2\right]$. *Next, we discuss a couple of special cases in Big-O notation with respect to the number of clients* m*, the number of local steps* s*, spectral norm* ρ*, stochastic gradient variance* σ *and bounded gradient dissimilarity* ζ*.*

- *FedPBC reduces to FedAvg with full-client participation when* $c = 1$*. Setting* $\eta = \sqrt{m/sT}$ *in Theorem 1, our convergence rate* $O(\frac{1}{\sqrt{msT}} + \sqrt{\frac{ms}{T}}\left(\sigma^2+\zeta^2\right))$ *matches the FedAvg literature (e.g., [30]).*
- *When it comes to FedAvg with uniform and time-invariant participation, suppose* k *out of* m *clients are selected uniformly at random each round. Setting* $\eta = \sqrt{k/sT}$ *in Theorem 1, our convergence rate becomes* $O(\frac{1}{\sqrt{ksT}} + \frac{1}{1-\sqrt{\rho}}\sqrt{\frac{ks}{T}}\left(\sigma^2+\zeta^2\right))$*. Since* $\rho \leq 1 - c^2/8$ *(in Lemma 3), the rate becomes* $O(\frac{1}{\sqrt{ksT}} + \frac{1}{c^2}\sqrt{\frac{ks}{T}}\left(\sigma^2+\zeta^2\right))$*, which introduces a larger variance compared to the rate of FedAvg with full participation, consistent with existing literature (e.g., [39]).*

(4) **On convergence rate.** *Our convergence rate in Corollary 1 of* $O(1/\sqrt{T})$*, where the Big-O notation is taken with respect to the total global round* T*, matches the best possible rate for any first-order algorithms that have access to only noisy stochastic gradients of a smooth non-convex objective [40]. By setting learning rate* η *as in bulletpoint (3), we shall see linear speedup with respect to the first term; however, the second term ultimately dominates the first term, which is consistent with FedAvg literature, see, e.g., [4]. We leave achieving linear speedup as a future direction.*

VII. Numerical Experiments

In this section, we evaluate FedPBC and multiple baseline algorithms on a simple quadratic function and real-world datasets.

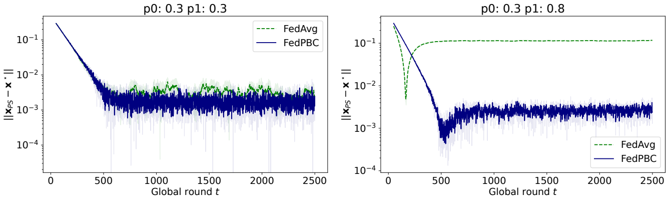


Fig. 3: $\|\boldsymbol{x}_{\text{PS}} - \boldsymbol{x}^\star\|_2$ in logarithmic scale. The results are obtained after an average of 3 random seeds. Plots are reported as mean ± standard deviation. The shaded areas plot standard deviation.

A. Quadratic function

The first part is about a simple quadratic function as in Eq. (2). Recall that, in each round t, client i responds to the parameter server's update request with probability p_i^t.

Counterexample. Our numerical results can be found in Fig. 3. We consider a federated learning system of $m = 100$ clients, each performing $s = 100$ steps local updates per round, in a total of 2500 global rounds. The local objective is $F_i(\boldsymbol{x}_i) = \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{u}_i\|_2^2$, where $\boldsymbol{x}_i, \boldsymbol{u}_i \in \mathbb{R}^{100}$, $\boldsymbol{u}_i \sim \mathcal{N}\left((i/1000)\mathbf{1}, 0.01\mathbf{I}\right)$, and $\boldsymbol{x}_i^0 = \mathbf{0}$ for all $i \in [m]$. The learning rate $\eta = 0.0001$. The uplinks of the first 50 clients become open with probability p_0, whereas the second half with p_1 – to be specified later. For ease of presentation, we plot the distance to the optimum $\|\boldsymbol{x}_{\text{PS}} - \boldsymbol{x}^\star\|_2$ after the first 50 communication rounds in Fig. 3, where $\boldsymbol{x}_{\text{PS}}^t \triangleq \boldsymbol{x}^t$ in Algorithm 1. All results are obtained after 3 random seeds and reported as mean ± standard deviation. Notably, all plots are on a logarithmic scale, potentially magnifying visual fluctuations. Notice that the distance to optimum $\|\boldsymbol{x}_{\text{PS}} - \boldsymbol{x}^\star\|_2$ does not go strictly to 0. We presumably attribute this to pseudo-randomness in computers to sample clients. Observe that two algorithms attain a similar distance to optimum when $p_0 = p_1$. Yet, FedPBC obtains a much lower error when $p_0 \neq p_1$. In addition, the error is on a similar scale (around 10^{-3}) as in the case of $p_0 = p_1$.

B. Real-world Datasets

In this section, we use three real-world datasets to validate the performance of FedPBC on different uplink unreliable patterns, and to compare with multiple baseline algorithms. Detailed hardware and software specifications can be found in Appendix B.

Dataset and data heterogeneity. The image classification task is commonly adopted in evaluating the empirical performance of a federated learning system [2], [20], [30], [34]. Following existing literature [2], [20], [30], [34], we base our simulations on SVHN [41], CIFAR-10 [42] and CINIC-10 [43]. All of them include 10 classes of images of different categories. For data heterogeneity, we partition all datasets and assign data samples to clients according to a Dirichlet distribution parameterized by α [44]. In particular, $\alpha = 0.1$ in Table. I. A smaller α entails a more non-i.i.d. local data distribution and vice versa. Each client holds the same data volume; the exact data volume may be dataset-dependent.

Federated learning system. We consider $m = 100$ clients, wherein clients continue to compute locally albeit the failures of unreliable communication uplinks. However, only clients with active links are allowed to submit their local updates. We use three customized convolutional neural networks for three datasets, respectively. Next, we introduce our construction of p_i^t's, which is then adopted to base the illustrations of unreliable patterns.

The construction of p_i^t's. We define

$$p_i^t \triangleq p_i \cdot \left[(1-\gamma) + \gamma \cdot \epsilon^t\right], \tag{14}$$

where $p_i \in (0, 1)$ is the time-invariant base probability, $\gamma \in [0, 1]$ is time-invariant and is used to control the variations of p_i^t, and ϵ^t is time-dependent. Detailed specifications are forthcoming.

- *Construction of* p_i. Inspired by [20], [27], the time-invariant base probability p_i is jointly determined by the local data distribution and a random variable R, which follows a $\mathsf{lognormal}(\mu_0, \sigma_0^2)$ distribution. It is immediately clear that the coupling leads to non-independent p_i's, which violates the assumption of independence in uplink communication failures in our theoretical analysis. However, FedPBC maintains its outperformance under such a challenging scenario. Define the number of classes in a dataset as C, the class distribution at a client i as $\boldsymbol{\nu}_i$ for $i \in [m]$. Since the local datasets are partitioned according to $\mathsf{Dirichlet}(\alpha)$, we have $\boldsymbol{\nu}_i \sim \mathsf{Dirichlet}(\alpha)$. Sample R from $\mathsf{lognormal}(\mu_0, \sigma_0^2)$ for C times to obtain a positive vector $\boldsymbol{r}' \in \mathbb{R}^C$. Normalize $\boldsymbol{r}'$ by dividing its l_1 norm and get $\boldsymbol{r} \triangleq \boldsymbol{r}'/\|\boldsymbol{r}'\|_1$. Finally, $p_i = \langle \boldsymbol{r}, \boldsymbol{\nu}_i \rangle$. Intuitively, $\boldsymbol{r}$ is used to quantify the unbalanced contribution of different classes. It is easy to see that for any fixed μ_0, a larger σ_0 leads to a more heterogeneous contribution distribution. We set $\mu_0 = 0$ and $\sigma_0 = 10$ in Table 1. By definition, p_i is a valid probability because
$$0 = \langle \mathbf{0}, \boldsymbol{\nu}_i \rangle < \langle \boldsymbol{r}, \boldsymbol{\nu}_i \rangle \overset{(a)}{\leq} \langle \boldsymbol{r}, \mathbf{1} \rangle = 1,$$
where $\mathbf{1}$ is an all-one vector, (a) holds because each element in $\boldsymbol{\nu}_i$ is no greater than 1, and p_i is strictly element-wise positive.
- *Construction of* ϵ^t. [7, Figure 5] indicates that the number of participants, i.e., clients with active communication uplinks, depends on time and acts like a sine curve. Inspired by this, we introduce a time-varying noise $\epsilon^t = \sin\left[(2\pi/P) \cdot t\right]$, where $P = 40$ defines the period and t is the current round index. This is a similar setup as the *Home Device* unreliable communication scheme in [12].
- *Choice of* γ. By definition, γ in (14) governs how severe the fluctuations of the sine curve in p_i^t's are. Given a fixed set of p_i's, γ determines both the lower and upper bounds of p_i^t's.

Fig. 4a presents an example of generated $\boldsymbol{r}$ drawn from a $\mathsf{lognormal}(0, 10^2)$, wherein class 0 and class 6 dominate the entire distribution. Intuitively, if a client i holds most of its images from classes other than 0 or 6, the generated p_i might be small and thus close to 0, possibly resulting in the client not appearing during training rounds in simulations. See Fig. 4b for details. To obtain meaningful results, we clip $p_i \leftarrow \max\{\delta, p_i\}$, where δ is a cutting-off parameter to ensure a lower bound on p_i. In Table I, $\delta = 0.02$. Notably, δ leads to

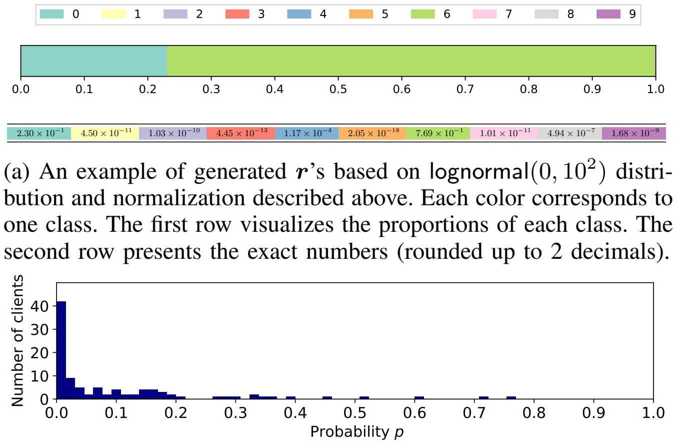
TABLE I: The reported results are in the form of mean accuracy ± standard deviation and are obtained over 3 repetitions in different random seeds. Results are averaged over the last 100 rounds. In each simulation, clients perform mini-batch stochastic gradient descent in 5 steps on a convolutional neural network (CNN) locally per round. The total global rounds for SVHN, CIFAR-10, CINIC-10 are 4000, 10000, 10000, respectively. Furthermore, we use customized CNNs for different datasets, respectively. Algorithms are categorized into two groups: (1) ones *not* aided by memory or known statistics; (2) ones with memory (including MIFA and FedAvg with *known* p_i^t's). Moreover, we highlight the best and the second best in yellow and in cyan, respectively, among algorithms *not* aided by memory or known statistics. The other hyperparameters are specified in Appendix, and some of them are tuned using grid search.

Unreliable Patterns	**Datasets**	SVHN		CIFAR-10		CINIC-10	
	Algorithms	Train	Test	Train	Test	Train	Test
Centralized		**88.7%**	**87.7%**	**76.1%**	**73.6%**	**61.9%**	**59.3%**
Bernoulli[1] with *time-invariant* p_i's	FedPBC (ours)	84.4% ± 0.008	84.3% ± 0.008	68.4% ± 0.011	66.3% ± 0.013	50.3% ± 0.005	49.7% ± 0.004
	FedAvg	75.9% ± 0.024	75.2% ± 0.024	59.9% ± 0.026	58.7% ± 0.025	38.1% ± 0.031	37.8% ± 0.029
	FedAvg *all*	56.4% ± 0.083	56.4% ± 0.072	48.9% ± 0.031	48.7% ± 0.026	32.6% ± 0.030	32.3% ± 0.030
	FedAU	83.1% ± 0.015	83.0% ± 0.015	67.4% ± 0.019	65.9% ± 0.019	45.8% ± 0.022	45.4% ± 0.022
	F3AST	76.9% ± 0.036	76.9% ± 0.037	58.5% ± 0.053	57.7% ± 0.052	40.7% ± 0.049	40.3% ± 0.048
	FedAvg *known* p_i's	77.8% ± 0.029	77.2% ± 0.032	61.1% ± 0.036	60.1% ± 0.035	39.2% ± 0.029	38.8% ± 0.029
	MIFA (*memory aided*)	80.8% ± 0.003	80.8% ± 0.003	67.8% ± 0.006	67.1% ± 0.006	47.6% ± 0.005	47.1% ± 0.005
Bernoulli with *time-varying* p_i^t's	FedPBC (ours)	84.0% ± 0.009	84.0% ± 0.009	67.1% ± 0.011	65.0% ± 0.015	49.7% ± 0.004	49.1% ± 0.003
	FedAvg	73.7% ± 0.041	72.7% ± 0.042	57.3% ± 0.034	56.2% ± 0.033	35.9% ± 0.038	35.6% ± 0.037
	FedAvg *all*	37.0% ± 0.097	36.5% ± 0.085	43.2% ± 0.030	43.2% ± 0.029	28.9% ± 0.024	28.7% ± 0.024
	FedAU	80.5% ± 0.023	80.3% ± 0.022	64.9% ± 0.018	63.5% ± 0.018	44.8% ± 0.017	43.4% ± 0.018
	F3AST	78.3% ± 0.027	78.1% ± 0.029	60.7% ± 0.037	59.6% ± 0.035	41.2% ± 0.035	40.8% ± 0.035
	FedAvg *known* p_i^t's	76.9% ± 0.035	76.3% ± 0.036	62.4% ± 0.021	61.2% ± 0.022	46.9% ± 0.016	46.4% ± 0.016
	MIFA (*memory aided*)	79.2% ± 0.005	79.2% ± 0.005	66.2% ± 0.006	65.5% ± 0.005	46.4% ± 0.010	45.8% ± 0.009
Homogeneous[1] Markovian with *time-invariant* p_i's	FedPBC (ours)	84.8% ± 0.009	84.1% ± 0.008	68.6% ± 0.010	66.5% ± 0.010	50.0% ± 0.006	49.5% ± 0.006
	FedAvg	74.7% ± 0.023	74.0% ± 0.023	59.1% ± 0.022	57.9% ± 0.020	37.4% ± 0.029	37.1% ± 0.029
	FedAvg *all*	55.1% ± 0.073	55.1% ± 0.063	48.3% ± 0.039	48.0% ± 0.034	31.6% ± 0.032	31.4% ± 0.031
	FedAU	82.7% ± 0.015	82.6% ± 0.013	68.3% ± 0.019	66.4% ± 0.018	47.2% ± 0.019	46.7% ± 0.018
	F3AST	75.5% ± 0.043	75.5% ± 0.048	60.3% ± 0.035	59.3% ± 0.034	43.0% ± 0.028	42.5% ± 0.027
	FedAvg *known* p_i's	76.0% ± 0.025	75.7% ± 0.027	61.0% ± 0.036	60.0% ± 0.034	40.8% ± 0.022	40.4% ± 0.022
	MIFA (*memory aided*)	81.7% ± 0.006	81.1% ± 0.004	66.8% ± 0.006	65.9% ± 0.006	46.9% ± 0.007	46.4% ± 0.007
Non-homogeneous Markovian with *time-varying* p_i^t's	FedPBC (ours)	83.9% ± 0.010	83.8% ± 0.008	67.2% ± 0.009	64.9% ± 0.006	49.7% ± 0.004	49.1% ± 0.004
	FedAvg	72.7% ± 0.034	72.2% ± 0.035	59.0% ± 0.027	58.0% ± 0.027	36.7% ± 0.031	36.3% ± 0.030
	FedAvg *all*	38.6% ± 0.091	38.3% ± 0.079	43.7% ± 0.026	43.8% ± 0.024	29.4% ± 0.025	29.2% ± 0.024
	FedAU	80.2% ± 0.020	80.2% ± 0.020	66.4% ± 0.018	65.1% ± 0.018	45.3% ± 0.022	44.8% ± 0.021
	F3AST	77.0% ± 0.033	77.0% ± 0.033	62.8% ± 0.032	61.5% ± 0.032	43.0% ± 0.029	42.6% ± 0.028
	FedAvg *known* p_i^t's	76.3% ± 0.045	76.3% ± 0.045	60.0% ± 0.040	59.0% ± 0.038	45.1% ± 0.032	44.5% ± 0.031
	MIFA (*memory aided*)	79.2% ± 0.005	79.1% ± 0.004	66.3% ± 0.007	65.6% ± 0.007	46.5% ± 0.008	46.1% ± 0.008
Cyclic[1] *without* periodic reset	FedPBC (ours)	84.2% ± 0.010	84.2% ± 0.009	67.5% ± 0.015	65.2% ± 0.017	49.7% ± 0.008	49.0% ± 0.007
	FedAvg	72.3% ± 0.029	71.7% ± 0.032	57.0% ± 0.028	56.0% ± 0.026	37.0% ± 0.029	36.6% ± 0.029
	FedAvg *all*	56.4% ± 0.078	56.4% ± 0.070	48.5% ± 0.026	48.1% ± 0.024	32.2% ± 0.028	31.9% ± 0.027
	FedAU	80.2% ± 0.027	79.8% ± 0.027	64.5% ± 0.024	63.1% ± 0.022	43.3% ± 0.033	42.8% ± 0.032
	F3AST	71.5% ± 0.042	71.7% ± 0.044	58.3% ± 0.026	57.3% ± 0.028	40.0% ± 0.028	39.7% ± 0.028
	FedAvg *known* p_i's[2]	74.1% ± 0.037	73.6% ± 0.038	58.9% ± 0.036	58.0% ± 0.034	38.1% ± 0.042	37.7% ± 0.041
	MIFA (*memory aided*)	70.9% ± 0.033	70.9% ± 0.033	59.1% ± 0.021	58.7% ± 0.022	42.3% ± 0.039	41.8% ± 0.038
Cyclic *with* periodic reset	FedPBC (ours)	83.8% ± 0.008	83.7% ± 0.007	66.3% ± 0.010	64.0% ± 0.012	49.6% ± 0.004	49.1% ± 0.004
	FedAvg	69.6% ± 0.054	69.0% ± 0.058	56.0% ± 0.032	55.1% ± 0.033	35.4% ± 0.027	35.1% ± 0.026
	FedAvg *all*	34.2% ± 0.074	33.6% ± 0.065	42.5% ± 0.026	42.4% ± 0.026	28.7% ± 0.023	28.5% ± 0.023
	FedAU	77.1% ± 0.029	77.1% ± 0.029	62.9% ± 0.022	61.7% ± 0.021	42.6% ± 0.020	42.1% ± 0.020
	F3AST	75.4% ± 0.035	75.3% ± 0.037	62.3% ± 0.041	61.0% ± 0.040	42.7% ± 0.041	42.2% ± 0.040
	FedAvg *known* p_i's[2]	72.7% ± 0.049	72.1% ± 0.052	60.0% ± 0.032	59.1% ± 0.030	45.5% ± 0.029	45.0% ± 0.028
	MIFA (*memory aided*)	77.6% ± 0.014	77.3% ± 0.014	64.8% ± 0.006	64.3% ± 0.006	45.6% ± 0.010	45.2% ± 0.010

the lower bound of p_i^t being $\delta \cdot (1 - 2\gamma)$. Now, we are ready to present unreliable schemes.

Unreliable schemes. In addition to a similar unreliable time-invariant communication setup as in [27] for fair competition, we study a more challenging scenario where p_i^t's change over time. Specifically, we evaluate FedPBC and a set of baseline algorithms on the following schemes:

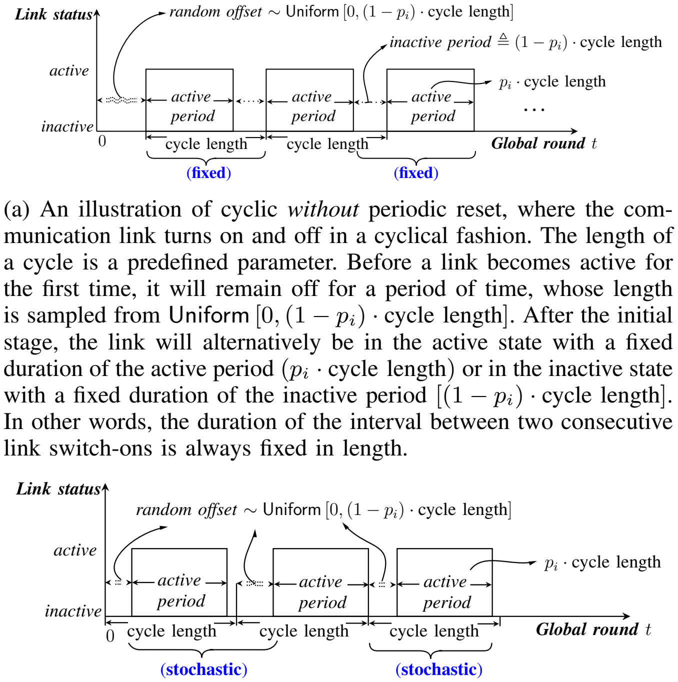
1) **Bernoulli.** Client i submits its local updates to the parameter server when the uplink becomes active with probability p_i^t. The first two columns of Table I demonstrate the results when the probabilities are *time-invariant* p_i's and *time-varying* p_i^t's, respectively. When p_i^t is time-invariant, we have $p_i^t = p_i$ for all $t \geq 0$, where p_i is the time-invariant base probability in (14). In the latter, p_i^t is defined as in (14) and changes over time.
2) **Markovian.** The uplink connection probabilities p_i^t's might be affected by external factors, leading to an unexpected shutdown after it is on or, conversely, resuming fully operational after it is off. Specifically, the uplink availability is dictated by a Markov chain of two states "ON" and "OFF", whose initial state is determined by a Bernoulli sampling. Depending on whether the transition probabilities change over time, we have a homogeneous Markov chain (the third row of Table I) or a non-homogeneous Markov chain (the fourth row). The detailed illustration of the transition probabilities is deferred to Appendix B.
3) **Cyclic.** The communication uplink between the parameter server and the clients can have a cyclic pattern, where the client has a fixed working schedule and joins the training diurnally or nocturnally [7], [23]. A random offset at the beginning of the whole process is used to simulate and reflect the initial shift due to each client's device



(a) An example of generated $\boldsymbol{r}$'s based on $\mathsf{lognormal}(0, 10^2)$ distribution and normalization described above. Each color corresponds to one class. The first row visualizes the proportions of each class. The second row presents the exact numbers (rounded up to 2 decimals).

(b) Histograms of the constructed p_i's under $R \sim \mathsf{lognormal}(0, 10^2)$ and $\boldsymbol{\nu}_i \sim \mathsf{Dirichlet}(0.1)$ with 100 clients and $\delta = 0$.

Fig. 4: The construction of p_i's.

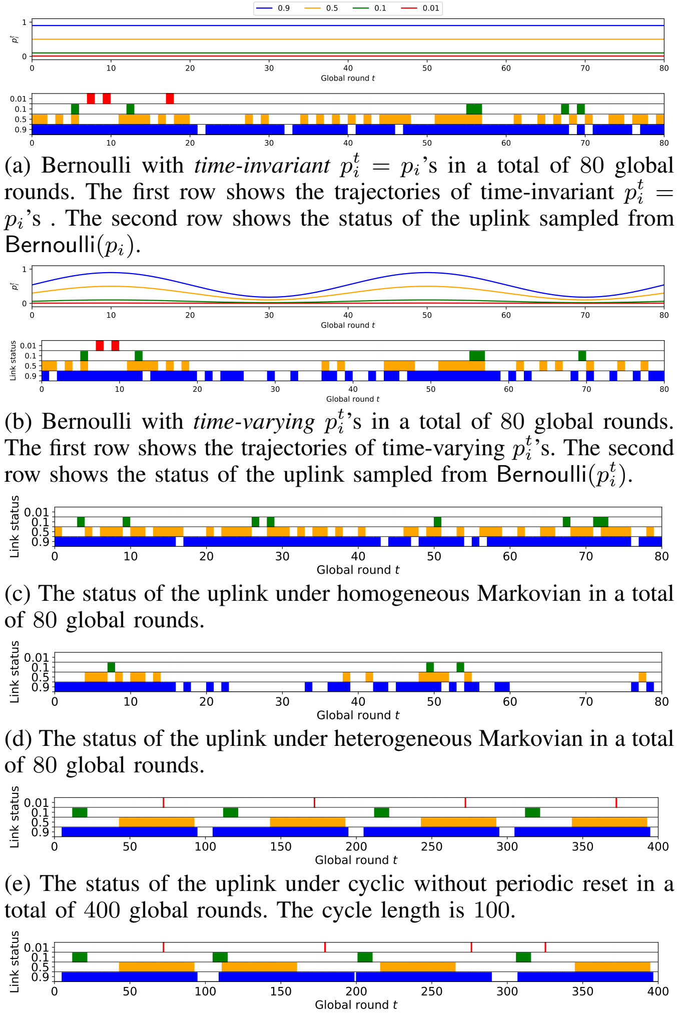


(a) An illustration of cyclic *without* periodic reset, where the communication link turns on and off in a cyclical fashion. The length of a cycle is a predefined parameter. Before a link becomes active for the first time, it will remain off for a period of time, whose length is sampled from Uniform $[0, (1-p_i) \cdot \text{cycle length}]$. After the initial stage, the link will alternatively be in the active state with a fixed duration of the active period $(p_i \cdot \text{cycle length})$ or in the inactive state with a fixed duration of the inactive period $[(1-p_i) \cdot \text{cycle length}]$. In other words, the duration of the interval between two consecutive link switch-ons is always fixed in length.

(b) An illustration of cyclic *with* periodic reset. Similar to Fig. 5a, a link switches on and off in alternation. The key difference is that a random offset will be redrawn from the same uniform distribution at the beginning of each cycle. The resampling procedure is called a reset, which entails a stochastic length of the interval between two consecutive link switch-ons.

Fig. 5: Illustrations of the communication unreliable schemes evaluated in Section VII-B

heterogeneity [27]. Please refer to Fig. 5a for details. However, it is also possible that each client's schedule to start training varies each day, which motivates us to devise the second scheme with periodic reset in Fig. 5b. The key difference is that the random offset will be reset at the beginning of each cycle, not only at the first cycle. Notice that the interval for a link to become active is now stochastic, rather than fixed.



(a) Bernoulli with *time-invariant* $p_i^t = p_i$'s in a total of 80 global rounds. The first row shows the trajectories of time-invariant $p_i^t = p_i$'s . The second row shows the status of the uplink sampled from $\mathsf{Bernoulli}(p_i)$.

(b) Bernoulli with *time-varying* p_i^t's in a total of 80 global rounds. The first row shows the trajectories of time-varying p_i^t's. The second row shows the status of the uplink sampled from $\mathsf{Bernoulli}(p_i^t)$.

(c) The status of the uplink under homogeneous Markovian in a total of 80 global rounds.

(d) The status of the uplink under heterogeneous Markovian in a total of 80 global rounds.

(e) The status of the uplink under cyclic without periodic reset in a total of 400 global rounds. The cycle length is 100.

(f) The status of the uplink under cyclic with periodic reset in a total of 400 global rounds. The cycle length is 100.

Fig. 6: Exemplary trajectories of p_i^t's and uplink status under different unreliable communication schemes. Colored blocks indicate that an uplink is active in the given round. We simulate the scenarios where $p_i \in \{0.01, 0.1, 0.5, 0.9\}$. The construction of p_i^t based on p_i can be found in Section VII-B.

Fig. 6 shows an example of uplink statuses under the unreliable communication schemes we evaluate. It is observed that uplinks become less frequently active when probabilities change from time-invariant (Fig. 6a) to time-varying (Fig. 6b). In addition, the uplinks become even more sparsely active when the schemes move to Markovian in Fig. 6c and 6d. On the other hand, the cyclic unreliable scheme exhibits a different pattern: the uplinks in Fig. 5a become active and inactive in alternation after an initial random offset. Notice that the uplink's offline duration is always fixed. In contrast, the duration remains random in Fig. 5b due to a reset at the beginning of each cycle.

Baseline algorithms. We compare FedPBC with six baseline algorithms, including FedAvg [2], FedAvg *all*, FedAvg *known* p_i^t*'s* [21], FedAU [27], F3AST [12], and MIFA [20]. Under FedAvg *all*, the parameter server averages all clients' local updates, wherein the contributions of clients with inactive

TABLE II: The first round to reach a targeted test accuracy under Bernoulli with *time-varying* p_i^t's over 3 random seeds. We study the first round to reach $1/4$, $1/2$, $3/4$ and 1 of the best test accuracy of each dataset in Table I, which is rounded up to the nearest 10% below for ease of presentation. In addition, we sample the mean of test accuracy every 150 global rounds to mitigate noisy progress. Some algorithms may never attain the targeted accuracy due to their inferior performance, where we use "–" as a placeholder. For example, the best test accuracy of FedAvg *all* is 36. 5% under Bernoulli with *time-varying* p_i^t's in Table I, below both 3/4 and 1 of the best accuracy.

Datasets	**Quarters**	**1/4**	**1/2**	**3/4**	**1**
	Test accuracy	20%	40%	60%	80%
SVHN	FedPBC (ours)	150	300	450	1650
	FedAvg	300	450	1050	–
	FedAvg *all*	1950	–	–	–
	FedAU	300	300	750	3450
	F3AST	450	750	1200	3600
	FedAvg *known* p_i^t's	600	1050	1650	–
	MIFA (*memory aided*)	300	600	1050	–
	Test accuracy	15%	30%	45%	60%
CIFAR-10	FedPBC (ours)	150	150	450	3300
	FedAvg	150	450	1050	9450
	FedAvg *all*	150	1500	–	–
	FedAU	150	300	750	3900
	F3AST	150	300	1200	4800
	FedAvg *known* p_i^t's	0	450	1800	4800
	MIFA (*memory aided*)	150	150	600	3600
	Test accuracy	10%	20%	30%	40%
CINIC-10	FedPBC (ours)	0	150	300	900
	FedAvg		150	1050	6450
	FedAvg *all*		600	–	–
	FedAU		150	300	2700
	F3AST		300	1200	3000
	FedAvg *known* p_i^t's	0	300	1050	2850
	MIFA (*memory aided*)		150	900	2700

communication links are deemed zeros. FedAvg *known* p_i^t*'s* requires the time-varying p_i^t's to be a known prior. We defer the other algorithmic specific parameters to Appendix B.

Results. Table I presents the evaluation results. The first row details the centralized learning results as a benchmark. We can see that all federated learning algorithms suffer some performance degradation, which is also commonly observed in distributed learning when there are communication constraints. Intuitively, this is the cost paid for not disclosing raw data to the other clients. In summary, FedPBC outperforms all other baseline algorithms *not* aided by memory on the SVHN and CINIC-10 datasets. In a rare instance, FedPBC is surpassed by FedAU on the CIFAR-10 dataset by a mere 0.2% in test accuracy. The rationale merits additional scrutiny. Additionally, FedAvg trails behind FedPBC by a substantial margin of approximately 10% in test accuracy, confirming its inherent bias.

It turns out that MIFA, aided by 100 units of old local gradients, does not always achieve the best performance. We conjecture it to the old gradients induced by a lower participation rate. Fig. 4b shows that most probabilities fall below 0.1 under our construction of p_i' s, which means that an uplink could be inactive for a long time before waking up again. Although clients in FedPBC start in each global round from its own staled local model, the expected staleness is upper bounded (see Proposition 2). It is not surprising that F3AST acts inferior to FedPBC. At a high level, F3AST caps $\mathcal{A}^t$ to a few representative clients for local optimization, excluding the rest of the clients within $\mathcal{A}^t$. FedPBC surpasses FedAU in all scenarios in terms of train accuracy. Although FedAU develops an online average method to estimate the underlying connection probabilities, it cannot tolerate complex dynamics. This can be observed in the performance degradation when switching from cyclic *without* periodic restart to cyclic *with* periodic restart. In the former, the uplinks are activated alternately with a fixed interval after the initial random offset, whereas in the latter, they are switched on stochastically, making it much more challenging. In the case of time-invariant p_i's, the outperformance of our FedPBC may stem from its utilization of true gradient trajectories to account for inactivities. This approach may result in better compensation than the online estimate used in FedAU. Though FedAvg with *known* probability uses the ground truth $1/p_i^t$ to mimic the empirical length of the uplink active interval, as pointed out in [27], the empirical length can unfortunately deviate far from the ground truth $1/p_i^t$.

To complement the numerical results in the main section, we also study the impact of different system-design parameters, including α, γ, δ, σ_0, on learning performance. The results are deferred to Appendix B.

Staleness. Table II demonstrates the first round to reach a targeted test accuracy under Benoulli with *time-varying* p_i^t's. Specifically, we study the round to reach the four quarters of the best test accuracy, which is rounded to the nearest 10% below for a neat presentation. It is readily seen that FedPBC attains a similar round to reach 1/4 and 1/2 of the best test accuracy as either FedAU or MIFA. When it is beyond 3/4 of the best accuracy, FedPBC in fact becomes the fastest algorithm. Hence, we empirically conclude that the staleness in FedPBC is mild and confirms its practicality.

VIII. Conclusion

In this paper, we study federated learning in the presence of stochastic uplink communications that are allowed to be simultaneously *time-varying* and *unknown* to all parties in the distributed learning system. We show that, by using a simple quadratic counterexample in Proposition 1, the seminal work FedAvg is inherently biased from the global optimum under non-i.i.d. local data. We propose FedPBC, which leverages implicit gossiping by postponing the broadcast till the end of each global round, is provable to reach a stationary point of the global non-convex objective, and converges at the optimal rate in the presence of smooth non-convex and stochastic objective gradients. Extensive experiments have been provided over diversified unreliable patterns to corroborate our analysis. Numerous directions are open for future research. First, FedPBC requires clients to perform local computation throughout training rounds, which may bring in extra computation costs. It is interesting to study how FedPBC can be applied to serve clients with limited computation resources. In addition, our work only addresses unreliable uplink communication. So, unreliable bidirectional communication failures are another extension. We expect to incorporate different local optimization methods, other than stochastic gradient descent, and establish provable guarantees. Finally, it is also interesting to explore achieving the desired linear speedup property.

References

[1] M. Xiang, S. Ioannidis, E. Yeh, C. Joe-Wong, and L. Su, "Towards bias correction of fedavg over nonuniform and time-varying communications," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, Dec 2023, pp. 6719–6724.

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

[3] 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," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[4] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJxNAnVtDS

[5] C. Philippenko and A. Dieuleveut, "Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees," *arXiv preprint arXiv:2006.14591*, 2020.

[6] S. Wang and M. Ji, "A unified analysis of federated learning with arbitrary client participation," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=qSs7C7c4G8D

[7] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.

[8] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.

[9] M. Zhang, M. Polese, M. Mezzavilla, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "Will tcp work in mmwave 5g cellular networks?" *IEEE Communications Magazine*, vol. 57, no. 1, pp. 65–71, 2019.

[10] Z. Guan and T. Kulkarni, "On the effects of mobility uncertainties on wireless communications between flying drones in the mmwave/thz bands," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 768–773.

[11] P. J. Mateo, C. Fiandrino, and J. Widmer, "Analysis of tcp performance in 5g mm-wave mobile networks," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–7.

[12] M. Ribero, H. Vikalo, and G. De Veciana, "Federated learning under intermittent client availability and time-varying communication constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 98–111, 2022.

[13] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[14] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," *Advances in neural information processing systems*, vol. 28, 2015.

[15] X. Zhang, J. Liu, and Z. Zhu, "Taming convergence for asynchronous stochastic gradient descent with unbounded delay in non-convex learning," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3580–3585.

[16] H. R. Feyzmahdavian, A. Aytekin, and M. Johansson, "An asynchronous mini-batch algorithm for regularized stochastic optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3740–3754, 2016.

[17] Y. Arjevani, O. Shamir, and N. Srebro, "A tight convergence analysis for stochastic gradient descent with delayed updates," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 111–132.

[18] S. U. Stich and S. P. Karimireddy, "The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 9613–9648, 2020.

[19] H. Yang, X. Zhang, P. Khanduri, and J. Liu, "Anarchic federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 331–25 363.

[20] X. Gu, K. Huang, J. Zhang, and L. Huang, "Fast federated learning in the presence of arbitrary device unavailability," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 052–12 064, 2021.

[21] J. Perazzone, S. Wang, M. Ji, and K. S. Chan, "Communication-efficient device scheduling for federated learning using stochastic optimization," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1449–1458.

[22] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 351–10 375.

[23] Y. J. Cho, P. Sharma, G. Joshi, Z. Xu, S. Kale, and T. Zhang, "On the convergence of federated averaging with cyclic client participation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5677–5721.

[24] W. Chen, S. Horváth, and P. Richtárik, "Optimal client sampling for federated learning," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=8GvRCWKHIL

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

[26] Y. Yan, C. Niu, Y. Ding, Z. Zheng, S. Tang, Q. Li, F. Wu, C. Lyu, Y. Feng, and G. Chen, "Federated optimization under intermittent client availability," *INFORMS Journal on Computing*, 2023.

[27] S. Wang and M. Ji, "A lightweight method for tackling unknown participation statistics in federated averaging," in *International Conference on Learning Representations*, 2024.

[28] D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi, "Fedvarp: Tackling the variance due to partial client participation in federated learning," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, J. Cussens and K. Zhang, Eds., vol. 180. PMLR, 01–05 Aug 2022, pp. 906–916. [Online]. Available: https://proceedings.mlr.press/v180/jhunjhunwala22a.html

[29] M. Tang and V. W. Wong, "Tackling system induced bias in federated learning: Stratification and convergence analysis," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[30] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[31] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* IEEE, 2003, pp. 482–491.

[32] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions," *Journal of Machine Learning Research*, vol. 21, no. 58, 2020.

[33] 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.

[34] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[35] X. Yuan and P. Li, "On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=_33ynl9VgCX

[36] L. Su, M. Xiang, J. Xu, and P. Yang, "Federated learning in the presence of adversarial client unavailability," *arXiv preprint arXiv:2305.19971*, 2024.

[37] J. Wang, A. K. Sahu, G. Joshi, and S. Kar, "Matcha: A matching-based link scheduling strategy to speed up distributed optimization," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5208–5221, 2022.

[38] M. Jerrum and A. Sinclair, "Conductance and the rapid mixing property for markov chains: the approximation of permanent resolved," in *Proceedings of the twentieth annual ACM symposium on Theory of computing*, 1988, pp. 235–244.

[39] H. Yang, M. Fang, and J. Liu, “Achieving linear speedup with partial worker participation in non-iid federated learning,” *arXiv preprint arXiv:2101.11203*, 2021.

[40] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[41] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” 2011.

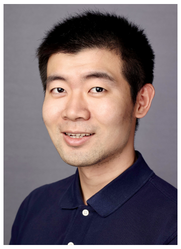
[42] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[43] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, “Cinic-10 is not imagenet or cifar-10,” *arXiv preprint arXiv:1810.03505*, 2018.

[44] T.-M. H. Hsu, H. Qi, and M. Brown, “Measuring the effects of non-identical data distribution for federated visual classification,” *arXiv preprint arXiv:1909.06335*, 2019.

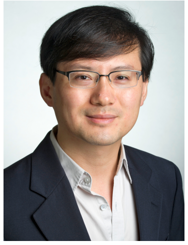
[45] M. Mandelbaum, M. Hlynka, and P. H. Brill, “Nonhomogeneous geometric distributions with relations to birth and death processes,” *Top*, vol. 15, pp. 281–296, 2007.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.



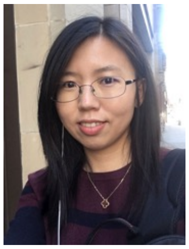
Ming Xiang (Student Member, IEEE) received the B.E. (2018) and M.E. (2021) in electrical engineering from Wuhan University, Wuhan, China. He is currently pursuing working towards the Ph.D. degree with Northeastern University in Boston, MA. His research interests include federated learning and fault-tolerance of distributed systems.

Stratis Ioannidis (Member, IEEE) received the B.Sc. degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 2002, and the M.Sc. and Ph.D. degrees in computer science from the University of Toronto, Canada, in 2004 and 2009, respectively. He is currently an Associate Professor with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA, USA, where he also holds a courtesy appointment with the Khoury College of Computer Sciences. Prior to joining Northeastern, he was a Research Scientist with the Technicolor Research Center, Paris, France, Palo Alto, CA, USA, and the Yahoo Laboratories, Sunnyvale, CA, USA. He was a recipient of the NSF CAREER Award, the Google Faculty Research Award, the Facebook Research Award, best paper awards at the 2017 ACM Conference on Information-Centric Networking (ICN), and the 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN).



Edmund Yeh (Senior Member, IEEE) received the B.S. degree (Hons.) in electrical engineering from Stanford University in 1994, the M.Phil. degree in engineering from Cambridge University in 1995 through the Winston Churchill Scholarship, and the Ph.D. degree in electrical engineering and computer science from MIT under Prof. Robert Gallager in 2001. He is currently a Professor in electrical and computer engineering at Northeastern University, with a Khoury College of Computer Sciences courtesy appointment.. Previously, he was an Assistant Professor and an Associate Professor in electrical engineering, computer science, and statistics with Yale University. He is an IEEE Communications Society Distinguished Lecturer. He was a recipient of the Alexander von Humboldt Research Fellowship, the Army Research Office Young Investigator Award, the Winston Churchill Scholarship, the National Science Foundation and Office of Naval Research Graduate Fellowships, the Barry M. Goldwater Scholarship, the Frederick Emmons Terman Engineering Scholastic Award, and the President’s Award for Academic Excellence (Stanford University). He has received three best paper awards, including awards from the 2017 ACM Conference on Information-Centric Networking (ICN) and the 2015 IEEE International Conference on Communications (ICC) Communication Theory Symposium. He serves as the TPC Co-Chair for ACM MobiHoc 2021. He also serves as a Treasurer of the Board of Governors for the IEEE Information Theory Society. He served as the General Chair for ACM SIGMETRICS 2020, an Associate Editor for IEEE TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, as the Guest Editor-in-Chief of the Special Issue on Wireless Networks for Internet Mathematics, and a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS—Special Series on Smart Grid Communications. He also received the Phi Beta Kappa Award.

Carlee Joe-Wong (Senior Member, IEEE) is the Robert E. Doherty Associate Professor of Electrical and Computer Engineering at Carnegie Mellon University. She received her A.B. degree (magna cum laude) in Mathematics, and M.A. and Ph.D. degrees in Applied and Computational Mathematics, from Princeton University in 2011, 2013, and 2016, respectively. Her research interests lie in optimizing various types of networked systems, including applications of machine learning and pricing to cloud computing, mobile/wireless networks, and ridesharing networks. From 2013 to 2014, she was the Director of Advanced Research at DataMi, a startup she co-founded from her research on mobile data pricing. She received the NSF CAREER award in 2018 and the ARO Young Investigator award in 2019.



Lili Su (Member, IEEE) is an Assistant Professor in the Electrical and Computer Engineering department at Northeastern University, in Boston, MA. She also holds a courtesy appointment with the Khoury College of Computer Sciences. She received her M.Sc. (2014) and Ph.D. (2017) in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign (UIUC). Prior to joining Northeastern, she was a postdoc in the Computer Science and Articial Intelligence Laboratory (CSAIL) at MIT. Her research intersects distributed systems - resilience and efficiency, distributed learning, federated learning, and multi-agent systems. She was the recipient of the NSF CAREER Award in 2024, runner-up of the Best Student Paper Award from the 30th International Symposium on Distributed Computing (DISC 2016), and she received the 2015 Best Student Paper Award from the 17th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2015). She received Sundaram Seshu International Student Fellowship from UIUC in 2016, and was selected as Rising Stars in EECS (2018).

Appendix A
Proofs

Proposition 3 is illustrated first as an intermediate result to assist in the proofs.

Proposition 3. *For any $t \in [T-1]$, it holds that*

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \le \frac{3L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2 + 3\zeta^2 \\ +3\left(\beta^2+1\right)\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2. \quad (15)$$

Inequality (15) can be shown by Jensen's inequality, where we plug in Assumptions 2 and 5.

Proof of Proposition 1. At each client $i \in \mathcal{A}^t$, for each local step $k = 0, \cdots, s-1$, we have

$$\boldsymbol{x}_i^{(t,k+1)} = (1-\eta)^{k+1}\boldsymbol{x}^t + \eta\boldsymbol{u}_i\left[\sum_{r=0}^{k}(1-\eta)^r\right].$$

It follows that

$$\begin{aligned}\boldsymbol{x}^{t+1} &= \mathbf{1}_{\{\mathcal{A}_t=\emptyset\}}\boldsymbol{x}^t \\ &\quad + \mathbf{1}_{\{\mathcal{A}_t\neq\emptyset\}}\frac{1}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t}\left((1-\eta)^s\boldsymbol{x}^t + \eta\boldsymbol{u}_i\left[\sum_{r=0}^{s-1}(1-\eta)^r\right]\right) \\ &= \boldsymbol{x}^t\mathbf{1}_{\{\mathcal{A}_t=\emptyset\}} + (1-\eta)^s\boldsymbol{x}^t\mathbf{1}_{\{\mathcal{A}^t\neq\emptyset\}} \\ &\quad + \frac{\eta\sum_{i\in\mathcal{A}^t}\boldsymbol{u}_i\left[\sum_{r=0}^{s-1}(1-\eta)^r\right]\mathbf{1}_{\{\mathcal{A}^t\neq\emptyset\}}}{|\mathcal{A}^t|} \\ &= \left[\mathbf{1}_{\{\mathcal{A}^t=\emptyset\}} + (1-\eta)^s\mathbf{1}_{\{\mathcal{A}^t\neq\emptyset\}}\right]\boldsymbol{x}^t \\ &\quad + \left[1-(1-\eta)^s\right]\frac{\mathbf{1}_{\{\mathcal{A}^t\neq\emptyset\}}}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t}\boldsymbol{u}_i.\end{aligned}$$

Taking expectation with respect to $\mathcal{A}^t$, we get

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}^{t+1}\mid\mathcal{A}^t\right] &= \left[\mathbb{P}\left\{\mathcal{A}^t=\emptyset\right\} + (1-\eta)^s\mathbb{P}\left\{\mathcal{A}^t\neq\emptyset\right\}\right]\boldsymbol{x}^t \\ &\quad + \left[1-(1-\eta)^s\right]\mathbb{E}\left[\frac{\sum_{i\in\mathcal{A}^t}\boldsymbol{u}_i}{|\mathcal{A}^t|}\Big|\mathcal{A}^t\neq\emptyset\right]\mathbb{P}\left\{\mathcal{A}^t\neq\emptyset\right\} \\ &= \left(\prod_{i=1}^{m}(1-p_i) + \left[1-\prod_{i=1}^{m}(1-p_i)\right](1-\eta)^s\right)\boldsymbol{x}^t \\ &\quad + \left[1-(1-\eta)^s\right]\left[1-\prod_{i=1}^{m}(1-p_i)\right]\mathbb{E}\left[\frac{\sum_{i\in\mathcal{A}^t}\boldsymbol{u}_i}{|\mathcal{A}^t|}\Big|\mathcal{A}^t\neq\emptyset\right].\end{aligned}$$

Following from the fact that $p_i^t = p_i$ for all t at all clients, $\mathbb{E}[\frac{1}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t}\boldsymbol{u}_i|\mathcal{A}^t\neq\emptyset] = \mathbb{E}[\frac{1}{|\mathcal{A}^1|}\sum_{i\in\mathcal{A}^1}\boldsymbol{u}_i|\mathcal{A}^1\neq\emptyset]$ for all t. Unrolling the above displayed equation until time 0 and applying the full expectation up to time $t+1$, we have

$$\mathbb{E}\left[\boldsymbol{x}^{t+1}\right] = \left(1-\mathrm{a}^{t+1}\right)\mathbb{E}\left[\mathbb{E}\left[\frac{1}{|\mathcal{A}^1|}\sum_{i\in\mathcal{A}^1}\boldsymbol{u}_i\Big|\mathcal{A}^1\neq\emptyset\right]\right], \quad (16)$$

where $\boldsymbol{x}^0 = \mathbf{0}$, and

$$\mathrm{a} \triangleq \prod_{i=1}^{m}(1-p_i) + \left[1-\prod_{i=1}^{m}(1-p_i)\right](1-\eta)^s.$$

Notably, $\mathrm{a} < 1$, it holds that $\lim_{t\to\infty}(1-\mathrm{a}^{t+1}) = 1$. Let $X_i = \mathbf{1}_{\{i\in\mathcal{A}^1\}}$ for each $i \in [m]$. We have

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{i\in\mathcal{A}^1}\boldsymbol{u}_i}{|\mathcal{A}^1|}\Big|\mathcal{A}^1\neq\emptyset\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^{m}X_i\boldsymbol{u}_i}{\sum_{j=1}^{m}X_j}\Big|\sum_{j=1}^{m}X_j\neq 0\right] \\ &= \sum_{i=1}^{m}\boldsymbol{u}_i\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\Big|\sum_{j=1}^{m}X_j\neq 0\right].\end{aligned}$$

By the law of total expectation and the convention that $\frac{0}{0} = 0$, we know that

$$\begin{aligned}\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\right] &= \mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\Big|\sum_{j=1}^{m}X_j\neq 0\right]\mathbb{P}\left\{\sum_{j=1}^{m}X_j\neq 0\right\} + 0 \\ &= \mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\Big|\sum_{j=1}^{m}X_j\neq 0\right]\mathbb{P}\left\{\sum_{j=1}^{m}X_j\neq 0\right\}.\end{aligned}$$

Hence,

$$\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{M}X_j}\Big|\sum_{j=1}^{M}X_j\neq 0\right] = \frac{\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\right]}{1-\prod_{i=1}^{m}(1-p_i)}.$$

Additionally,

$$\begin{aligned}\mathbb{E}\left[\frac{X_i}{\sum_{i=1}^{m}X_i}\right] &= \mathbb{P}\left\{X_i=1\right\}\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\Big|X_i=1\right] + 0 \\ &= p_i\mathbb{E}\left[\frac{1}{1+\sum_{j\in[m]\setminus\{i\}}X_j}\Big|X_i=1\right]. \quad (17)\end{aligned}$$

Next, we show that

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{1+\sum_{j\in[m]\setminus\{i\}}X_j}\Big|X_i=1\right] \\ &\qquad = 1 + \sum_{j=2}^{m}(-1)^{j+1}\frac{1}{j}\sum_{S\in\mathcal{B}_j^i}\prod_{z\in S}p_z, \quad (18)\end{aligned}$$

where $\mathcal{B}_j^i \triangleq \left\{S\Big|S\subseteq[m]\setminus\{i\}, |S| = j-1\right\}$. Without loss of generality, assume $i = m$. Define $\bar{S} \triangleq [m]\setminus S$

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{1+\sum_{j\in[m]\setminus\{m\}}X_j}\Big|X_m=1\right] = \mathbb{E}\left[\frac{1}{1+\sum_{j\in[m-1]}X_j}\right] \\ &\triangleq \sum_{j=1}^{m}\frac{1}{j}\mathbb{P}\left\{\left|\mathcal{A}^1\setminus\{m\}\right| = j-1\right\} \\ &= \sum_{j=1}^{m}\frac{1}{j}\sum_{S\in\mathcal{B}_j}\prod_{x\in\bar{S}}(1-p_x)\prod_{z\in S}p_z. \quad (19)\end{aligned}$$

Then, we show that (18) and (19) are equivalent. The degree coefficient of polynomial 0 (i.e., when $|S| = 0$) relates only to $j \in \{1\}$: $\prod_{k=1}^{m-1}(1-p_k)$, where we select all the ones in parentheses. Thus, the coefficient of the terms in the degree of polynomial 0 is 1. The degree coefficient of polynomial 1 (i.e., when $|S|=1$). corresponds to $j\in\{1,2\}$:

$$\prod_{k=1}^{m-1}(1-p_k)\ (j=1); \tag{20}$$

$$\frac{1}{2}\sum_{k=1}^{m-1}p_k\prod_{x\in[m-1]\setminus\{k\}}(1-p_x)\ (j=2). \tag{21}$$

Take the coefficient of p_1 as an example. In (20), to get p_1, we select p_1 from $(1-p_1)$ and all the ones from the rest parentheses, which yields $-1\binom{1}{0}$. In addition, in (21), the coefficient is $\frac{1}{2}\binom{1}{1}$. They add up to $-1+\frac{1}{2}=-\frac{1}{2}$. For a general degree coefficient of polynomial K (i.e., when $|S|=K$), by using a similar argument, the coefficient is $(-1)^K\left[\sum_{y=0}^{K}\frac{(-1)^y}{y+1}\binom{K}{y}\right]$, which can be simplified as

$$\begin{aligned}
&(-1)^K\sum_{y=0}^{K}\frac{(-1)^y}{y+1}\binom{K}{y}=(-1)^K\sum_{y=0}^{K}\frac{(-1)^y}{y+1}\frac{K!}{y!(K-y)!}\\
&=(-1)^K\frac{1}{K+1}\sum_{y=0}^{K}(-1)^y\frac{(K+1)!}{(y+1)!(K-y)!}\\
&=\frac{(-1)^{K+1}}{K+1}\sum_{y=0}^{K}(-1)^{y+1}\binom{K+1}{y+1}\\
&=\frac{(-1)^{K+1}}{K+1}\left[(-1+1)^{K+1}-(-1)^0\right]=\frac{(-1)^K}{K+1}.
\end{aligned}$$

Combining the above yields (18). Finally, we plug Eq. (17) in Eq. (16) and get

$$\begin{aligned}
&\lim_{t\to\infty}\mathbb{E}\left[\boldsymbol{x}^{t+1}\right]=\lim_{t\to\infty}\mathbb{E}\left[\sum_{i=1}^{m}\boldsymbol{u}_i\mathbb{E}\left[\frac{X_i}{\sum_{j=1}^{m}X_j}\Big|\mathcal{A}^1\neq\emptyset\right]\right]\\
&=\sum_{i=1}^{m}\frac{\boldsymbol{u}_ip_i\left(1+\sum_{j=2}^{m}(-1)^{j+1}\frac{1}{j}\sum_{S\in\mathcal{B}_j}\prod_{z\in S}p_z\right)}{1-\prod_{i=1}^{m}(1-p_i)},
\end{aligned}$$

where $\mathcal{B}_j\triangleq\left\{S\Big|S\subseteq[m]\setminus\{i\},|S|=j-1\right\}$.

Special cases.

a) When probabilities are uniform, i.e., $p_i=p$ for $i\in[m]$: The coefficient of each term in (3) becomes

$$\begin{aligned}
&\frac{p(1+\sum_{j=2}^{m}(-1)^{j+1}\frac{\binom{m-1}{j-1}}{j}p^{j-1})}{1-(1-p)^m}\overset{(a)}{=}\frac{p(1+\sum_{j=2}^{m}(-1)^{j+1}\frac{\binom{m}{j}}{m}p^{j-1})}{1-(1-p)^m}\\
&=\frac{1}{m}\cdot\frac{mp+\sum_{j=2}^{m}(-1)^{j+1}\binom{m}{j}p^j}{1-(1-p)^m}\\
&\overset{(b)}{=}\frac{1}{m}\cdot\frac{mp+\sum_{j=2}^{m}(-1)^{j+1}\binom{m}{j}p^j}{mp+\sum_{j=2}^{m}(-1)^{j+1}\binom{m}{j}p^j}=\frac{1}{m},
\end{aligned}$$

where equality (a) holds because $j\binom{m}{j}=m\binom{m-1}{j-1}$, equality (b) holds because

$$1-(1-p)^m\overset{(c)}{=}\sum_{j=1}^{m}(-1)^{j+1}\binom{m}{j}p^j=mp+\sum_{j=2}^{m}(-1)^{j+1}\binom{m}{j}p^j,$$

where equality (c) holds because of binomial theorem. Consequently, (3) reduces to the unbiased global optimum

$$\lim_{T\to\infty}\mathbb{E}\left[\boldsymbol{x}^T\right]=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{u}_i=\boldsymbol{x}^\star.$$

b) When clients local distributions are homogeneous, e.g., $\boldsymbol{u}_i=\boldsymbol{u}$ for all $i\in[m]$: (3) reduces to

$$\frac{\sum_{i=1}^{m}p_i\left[1+\sum_{j=2}^{m}(-1)^{j+1}\frac{1}{j}\sum_{S\in\mathcal{B}_j}\prod_{z\in S}p_z\right]}{1-\prod_{i=1}^{m}(1-p_i)}\boldsymbol{u}. \tag{22}$$

Let us define $\mathcal{C}_j\triangleq\left\{S'\Big|S'\subseteq[m],|S'|=j\right\}$. Next, we show that $\sum_{S'\in\mathcal{C}_j}\prod_{z'\in S'}p_{z'}=\sum_{i=1}^{m}\frac{p_i}{j}\sum_{S\in\mathcal{B}_j}\prod_{z\in S}p_z$. We start from the R.H.S. Take the occurrence of $A=p_{x_1}p_{x_2}\ldots p_{x_j}$ as an example, where $x_1<x_2<\ldots<x_j$. Since it is equally possible for p_{x_1}, p_{x_2}, $\ldots$ and p_{x_j} to be the leading term (i.e., p_i in (22)), we then have $\binom{j}{1}$ many A terms in the R.H.S. $\binom{j}{1}=j$ will cancel the original coefficient $\frac{1}{j}$ at each term. Hence, the equality holds. Consequently, (22) simplifies to (23).

$$\frac{\sum_{j=1}^{m}(-1)^{j+1}\sum_{S'\in\mathcal{C}_j}\prod_{z'\in S'}p_{z'}}{1-\prod_{i=1}^{m}(1-p_i)}=1, \tag{23}$$

where the equality holds because of the expansion of the term $\prod_{i=1}^{m}(1-p_i)$. Finally, we get

$$\lim_{T\to\infty}\mathbb{E}\left[\boldsymbol{x}^T\right]=\boldsymbol{u}. \tag{24}$$

(24) indicates that the global objective will recover each client's local optimums under even heterogeneous participation probability p_i's when clients' local data distributions are homogeneous. □

Proof of Proposition 2. In our work, the probabilities $p_i^t\geq c$. Therefore, define $Y_{\min}$ as the random variable of the ordinary geometric distribution with success probability c. We have $\mathbb{E}\left[Y_{\min}\right]=1/c$. [45, Theorem 3.2] tells us that $\mathbb{E}\left[t-\tau_i(t)\right]\leq\mathbb{E}\left[Y_{\min}\right]=1/c$. □

Proof of Theorem 1. In this proof, we combine all the above intermediate results to show the final theorem.

(a) Taking expectation over the remaining randomness and a telescoping sum.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1})-F(\bar{\boldsymbol{x}}^t)\right]\leq-\frac{s\eta}{3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right]\\
&\quad+6L\eta^2s^2\left(\kappa^2L^2+1\right)\left(\sigma^2+\zeta^2\right)+\eta s\frac{L^2}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right],
\end{aligned}$$

where inequality (a) holds because of Assumption 4.

(b) Plugging in Lemma 4 and Assumption 4.

$$\begin{aligned}
&\frac{F^\star-\mathbb{E}\left[F(\bar{\boldsymbol{x}}^0)\right]}{T}\\
&\leq 9\eta^2s^2L\left[\kappa^2L^2+1+16\eta s^2\frac{\rho sL}{\left(1-\sqrt{\rho}\right)^2}\right]\left(\sigma^2+\zeta^2\right)\\
&\quad-\frac{s\eta}{3}\left(1-162\eta^2s^2\frac{\rho\left(\beta^2+1\right)L^4}{\left(1-\sqrt{\rho}\right)^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right].
\end{aligned} \tag{25}$$

We know from $\eta \le \frac{1-\sqrt{\rho}}{108L^2s^3(\beta^2+1)(1+\kappa^2L^2)} \le \frac{1-\sqrt{\rho}}{18(\beta^2+1)L^2s}$ that

$$1 - 162\eta^2 s^2 \frac{\rho\left(\beta^2+1\right)L^4}{\left(1-\sqrt{\rho}\right)^2} \ge 1 - \frac{162\rho\left(\beta^2+1\right)L^4}{\left(1-\sqrt{\rho}\right)^2}\frac{\left(1-\sqrt{\rho}\right)^2}{324\left(\beta^2+1\right)^2L^4} \ge \frac{1}{2}.$$

In addition, we also have $\kappa^2L^2+1+16\eta s^3\frac{\rho L}{\left(1-\sqrt{\rho}\right)^2} \le \kappa^2L^2+1+\frac{1}{1-\sqrt{\rho}}$. Therefore, rearrange the terms in (25), it follows that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \le \frac{6\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta s T} + 54\eta s L\left(\kappa^2L^2+1+\frac{1}{1-\sqrt{\rho}}\right)\left(\sigma^2+\zeta^2\right).$$

□

Appendix B
Experimental Setup

Hardware and Software Setups. The simulations are performed on a private cluster with 64 CPUs, 500 GB RAM and 8 NVIDIA A5000 GPU cards. We code the experiments based on PyTorch 1.13.1 [46] and Python 3.7.16. Our code is accessible at `https://github.com/mingxiang12/FedPBC`.

Neural Network and Hyper-parameter Specifications. We initialize the customized CNNs using the Kaiming initialization. A decaying learning rate schedule $\eta = \eta_0/\sqrt{(t/10)+1}$ is adopted. The initial local learning rate η_0 and the global learning rate η_g are searched, based on the best performance after 500 global rounds, over two grids $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$ and $\{0.5, 1, 1.5, 5, 10, 50\}$, respectively. We set $\beta = 0.01$, which is tuned over a grid of $\{1, 0.5, 0.1, 0.05, 0.01, 0.005\} \times 10^{-2}$, for F3AST [12].

Missing algorithm descriptions. In this section, we specify the missing essential hyperparameters for specific algorithm implementations. As recommended by [27], we choose $K = 50$ for FedAU without further specification. Note that K is an algorithmic hyperparameter in FedAU. Adopting the setup in [12], we set the communication constraint to be 10 clients for F3AST.

Datasets. All the datasets we evaluate contain 10 classes of images. Some data enhancement tricks that are standard in training image classifiers are applied during training. Specifically, we apply random cropping to all datasets. Furthermore, random horizontal flipping is applied to CIFAR-10 and CINIC-10. SVHN [41] dataset contains 32×32 colored images of 10 different number digits. In total, there are 73257 train images and 26032 test images. CIFAR-10 [42] dataset contains 32×32 colored images of 10 different objects. In total, there are 50000 train images and 10000 test images. CINIC-10 [43] dataset contains 32×32 colored images of 10 different objects. In total, there are 90000 train images and 90000 test images.

Transition probabilities Conditions	$q_i^{t\star}$	q_i^t
$q_i^{t\star}\cdot(1-p_i^t) \le p_i^t$	0.05	$0.05\cdot\frac{1-p_i^t}{p_i^t}$
$q_i^{t\star}\cdot(1-p_i^t) > p_i^t$	$\frac{p_i^t}{1-p_i^t}$	1

TABLE III: The construction of q_i^t and $q_i^{t\star}$.

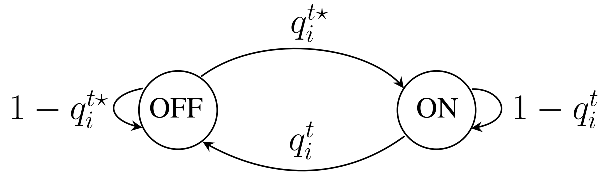


Fig. 7: An illustration of the Markovian transition probabilities.

Constructions of Markov transition probabilities. Recall that the link status in Markovian unreliable scheme is dictated by a Markov chain, whose initial states are based on Bernoulli(p_i^t). Fig. 7 plots the Markov chain. Let q_i^t and $q_i^{t\star}$ define the transition probability from the "ON" state to the "OFF" state and from the "OFF" state to the "ON" state, respectively. In the experiments, we aim to construct q_i^t and $q_i^{t\star}$ so that a stationary distribution is met as

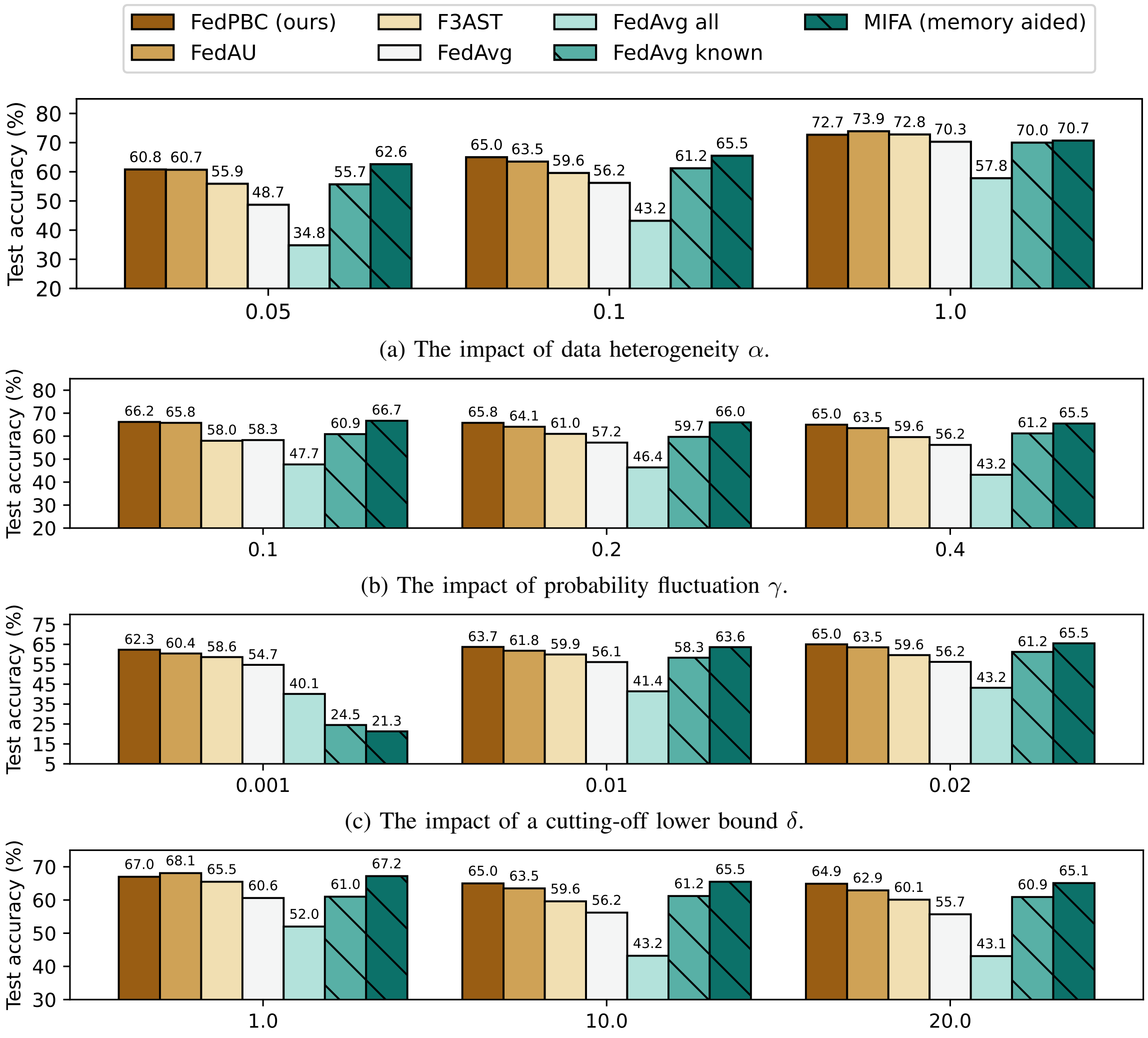
$$q_i^t \cdot p_i^t = q_i^{t\star} \cdot \left(1-p_i^t\right). \tag{26}$$

Concretely, we first assume that $q_i^{t\star} = 0.05$ is an external choice. If $q_i^{t\star}\cdot(1-p_i^t) > p_i^t$, we adjust q_i^t and $q_i^{t\star}$ to ensure (26). Please find the details in Table III.

Ablation Experiments. In this part, we conduct ablation experiments to study the impact of different parameters on the performance of FedPBC and the other baseline algorithms. Specifically, we evaluate all algorithms on the SVHN dataset under the Bernoulli unreliable communication scheme with *time-varying* p_i^t's. In any set of experiments, only one system design parameter is changed, while the others remain the same as in Table I. We report the mean test accuracy over the last 100 rounds in bar plots in Fig. 8. Algorithms are divided into two groups: those with additional memory or *known* historical statistics (bars with backslashes) and those without. It is observed that FedPBC outperforms the baseline algorithms *not* aided by memory in almost all cases (except when $\alpha = 1.0$ by FedAU in Fig. 8a and $\sigma_0 = 1.0$ by FedAU in Fig. 8d.) The reason why FedPBC trails behind FedAU in the above two cases is worth further investigation. Compared to memory-aided algorithms, although MIFA occasionally dwarfs FedPBC, the benefit margin is lower than 2% in test accuracy.

Impact of data heterogeneity α. In the presence of more homogenous local data, i.e., a larger α, the bias phenomenon gradually disappears as the local objectives become interchangeable, which is confirmed by Fig. 8a from the on-par performance of almost all algorithms when $\alpha = 1.0$.

Impact of fluctuation γ. The magnitude of the sine function is defined as γ and thus governs the fluctuations of p_i^t's. It can be seen that the test accuracies of all algorithms decrease as γ increases. This is intuitive, as enlarged fluctuations impose new challenges. It is observed that FedPBC outperforms all algorithms that are *not* aided by memory.



(a) The impact of data heterogeneity α.

(b) The impact of probability fluctuation γ.

(c) The impact of a cutting-off lower bound δ.

(d) The impact of contribution heterogeneity σ_0.

Fig. 8: The test accuracies in the ablation experiments. In each plot, only one system design parameter is changed. The others remain the same as in Table I. All experiments are evaluated on the SVHN dataset under Bernoulli with *time-varying* unreliable uplinks. The bars with backslashes refer to the algorithms requiring extra memory or *known* historical statistics.

Impact of a cutting-off lower bound δ. Recall that p_i's might be too small and close to 0 due to the unbalanced class contributions in $\boldsymbol{r}$. We show in Lemma 3 that a smaller lower bound c of p_i^t's slows down convergence and incurs a looser bound in Theorem 1. Notice that FedPBC remains the best among the algorithms *not* aided by memory in terms of test accuracy. At one challenging extreme (when $\delta = 0.001$), all algorithms experience significant drops in accuracy, in particular MIFA. This confirms our conjecture that the old gradient might lead to staled updates and affect performance.

Impact of contribution heterogeneity σ_0. A smaller σ_0 leads to a more even contribution of each class and thus more homogeneous p_i's. Hence, it is not surprising to find that many baseline algorithms attain accurate test predictions when $\sigma_0 = 1.0$. In contrast, FedPBC shadows all baseline algorithms except MIFA in the highly heterogeneous scenario where $\sigma_0 = 20.0$.