2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Towards Bias Correction of FedAvg over Nonuniform and Time-Varying Communications

Ming Xiang*, Stratis Ioannidis*, Edmund Yeh*, Carlee Joe-Wong†, and Lili Su*

Abstract— **Federated learning (FL) is a decentralized learning framework wherein a parameter server (PS) and a collection of clients collaboratively trains a model via minimizing a global objective. Communication bandwidth is a scarce resource; in each round, the PS aggregates the updates from a subset of clients only. In this paper, we focus on non-convex minimization that is vulnerable to non-uniform and time-varying communication failures between the PS and the clients. Specifically, in each round t, the link between the PS and client i is active with probability p_i^t, which is *unknown* to both the PS and the clients. This arises when the channel conditions are heterogeneous across clients and are changing over time.**

We show that when the p_i^t's are not uniform, *Federated Average* (FedAvg) – the most widely adopted FL algorithm – fails to minimize the global objective. Observing this, we propose *Federated Postponed Broadcast* (FedPBC) which is a simple variant of FedAvg. It differs from FedAvg in that the PS postpones broadcasting the global model till the end of each round. We show that FedPBC converges to a stationary point of the original objective. The introduced staleness is mild and there is no noticeable slowdown. Both theoretical analysis and numerical results are provided. On the technical front, postponing the global model broadcasts enables implicit gossiping among the clients with active links at round t. Despite p_i^t's are time-varying, we are able to bound the perturbation of the global model dynamics via the techniques of controlling the gossip-type information mixing errors.

I. INTRODUCTION

Federated learning (FL) is a distributed learning paradigm wherein a parameter server (PS) and a large collection of clients collaboratively learn a machine learning model with clients' local data undisclosed [1], [2] to the PS. The global objetives are often non-convex. Communication bandwidth is a scarce resource. In each round, the PS aggregates the updates from a subset of clients only – either proactively [1], [2] or passively [3]–[5]. A FL system is often deployed in a uncontrolled environment, wherein the channel conditions between the PS and the clients could be highly heterogeneous and time-varying [1]. To capture this, in this paper, we consider non-convex minimization that is vulnerable to non-uniform and time-varying link failures between the PS and the clients. Specifically, in each round, the link between the PS and client i is active with probability p_i^t, which is unknown to both the PS and the clients. A generic FL system of interest is illustrated in Fig. 1. To the best of our knowledge, the convergence of FL in the presence of non-uniform and time-varying communication is overall under-explored.

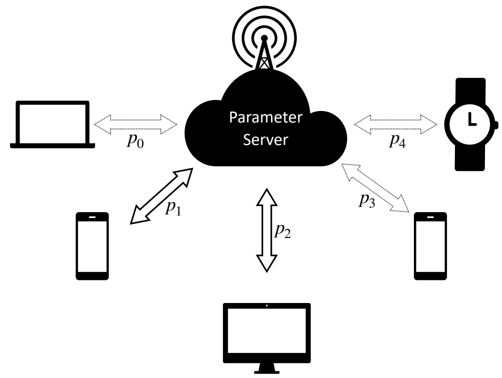


Fig. 1: A federated learning system with heterogeneous devices: Solid arrows indicate active links and dashed arrows are inactive links.

Our setup can be viewed as a special case of the general client unavailability, has received intensive attention recently [2]. Nevertheless, existing methods are not applicable to our problem. In the seminal works [1], [3], the PS chooses K clients either uniformly at random or proportionally to clients' local data volume. Neither of theses client selection methods is feasible when p_i^t's are unknown and time-varying. In [2]–[4], [6], the PS waits for the K fastest responses. The correctness of their algorithms crucially relies on the fact that the response probability of each client is known. Ruan et al. [7] considered a generalized random client unavailability, yet required the response probability to be fixed. Time-varying response rates are also considered in [5], [8], [9]. For the methods in [5] to converge to stationary points, the response rates need to be "balanced" in the sense that either (1) the p_i^t's are deterministic and satisfy the regularized participation, i.e., $\sum_{\tau=1}^{P} p_i^{t_0+\tau} = \mu$ 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 t. In contrast, we do not require such rate "balanceness". Perazzone et al. [8] analyzed the convergence of FedAvg under time-varying client participation rates. Nevertheless, they assumed (1) a uniform participation rate in each round, i.e., $p_i^t = p_j^t$ for

* Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02215, USA xiang.mi@northeastern.edu, {ioannidis,eyeh}@ece.neu.edu, l.su@northeastern.edu.
† Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA cjoewong@andrew.cmu.edu.
The work was supported in part by ARO Grant W911NF-23-2-0014 and in part by National Science Foundation under Grants 2107062 and 2106891.

any pair of clients, and (2) bounded stochastic gradient. Gu et al. [9] considered general client unavailability patterns for both strongly convex and non-convex global objectives. For non-convex objectives (which is our focus), they required 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 required the noise of the stochastic gradient to be uniformly upper bounded with probability 1.

Contributions. Our contributions is three-fold:

- We identify simple instances and show both analytically and numerically that when the p_i^t's are not uniform *Federated Average* (FedAvg) – the most widely adopted FL algorithm – fails to minimize the global objective.
- We propose *Federated Postponed Broadcast* (FedPBC). It differs from FedAvg in that the PS postpones broadcasting the global model till the end of each round. We show in Theorem 1 that, in expectation, FedPBC converges to a stationary point of the global objective. The correctness of our FedPBC *neither* impose any "balancedness" requirement on p_i^t's *nor* require the stochastic gradients or their noises to be bounded. Moreover, compared with [5], [9], FedPBC works under a much relaxed bounded-dissimilarity assumption.
  On the technical front, postponing the global model broadcasts enables implicit gossiping among the clients with active links. Hence, we mitigate the perturbation caused by non-uniform and time-varying p_i^t via the techniques of controlling information mixing errors.
- We validate our results empirically both on the counterexample and by using Synthetic $(1, 1)$ dataset [10]. The numerical results in the former show that FedPBC successfully corrects the bias when p_i^t's are static but non-uniform (i.e., $p_i^t = p_i$) while FedAvg does not. In the latter, we further investigate *time-varying* link activation rates such that the responsive rates follow a uniform distribution and thus are bounded below. The results show FedPBC outperforms FedAvg.

II. Problem Formulation

A FL system consists of one central PS and m clients that collaboratively minimize

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

where $F_i\left(\boldsymbol{x}\right) = \mathbb{E}_{\xi_i\in\mathcal{D}_i}\left[\ell_i\left(\boldsymbol{x};\xi_i\right)\right]$ 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 links between the PS and the clients. In each round t, the communication link between the PS and client i is active with probability p_i^t, which could be **time-varying** and is **unknown** to both the PS and the clients. We assume that $p_i(t) \ge c$ for all t and all i, where $c \in (0, 1)$.

III. A Case Study on the Objective Inconsistency of FedAvg

In this section, we use a simple example (a similar setup as in [11]) to illustrate FedAvg fails to minimize the global objective in Eq. (1) when p_i's are not uniform. For completeness, we formally describe FedAvg in Algorithm 1. Notably, in Algorithm 1, all the clients (regardless of whether

Algorithm 1: Federated Average (FedAvg) [1]

1 **Input:** T, $\boldsymbol{x}^0$, s, $\{\eta_t\}_{t=0,\cdots,T-1}$
2 The PS and each client initialize parameter $\boldsymbol{x}^0$;
3 **for** $t = 0, \cdots, T-1$ **do**
  /* Let $\mathcal{A}^t$ denote all the clients with active communication links. */
4   The PS broadcasts $\boldsymbol{x}^t$ to each client;
5   **for** $i \in [m]$ **do**
6   Draw a fresh sample ξ_i^t;
7   **if** $i \in \mathcal{A}^t$ **then**
8       $\boldsymbol{x}_i^{(t,0)} \leftarrow \boldsymbol{x}^t$;
9   **else**
10       $\boldsymbol{x}_i^{(t,0)} \leftarrow \boldsymbol{x}_i^t$;
11   **end**
12   **for** $k = 0, \cdots, s-1$ **do**
13       $\boldsymbol{x}_i^{(t,k+1)} \leftarrow \boldsymbol{x}_i^{(t,k)} - \eta_t \nabla \ell_i(\boldsymbol{x}_i^{(t,k)}; \xi_i^t)$;
14   **end**
15     $\boldsymbol{x}_i^{t++} \leftarrow \boldsymbol{x}_i^{(t,s)}$;
16   Report $\boldsymbol{x}_i^{t++}$ to the PS;
17   **end**
  /* On the PS. */
18   **if** $\mathcal{A}^t \neq \emptyset$ **then**
19     $\boldsymbol{x}^{t+1} \leftarrow \frac{1}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t} \boldsymbol{x}_i^{t++}$;
20   **else**
21     $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t$;
22   **end**
23 **end**

the corresponding links are active or not) compute locally in Algorithm 1 in each round. This is *logically equivalent* to the usual setting where only clients in $\mathcal{A}^t$ do the local steps because in line 20 the summation is taken over the clients in $\mathcal{A}^t$. Similar equivalence is observed in [5]. We present the FedAvg in the form of Algorithm 1 for ease of comparison with our FedPBC – an algorithmic fix to FedAvg for bias correction.

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

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

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

Proposition 1. *Choose $\boldsymbol{x}^0 = \boldsymbol{0}$ and $\eta_t = \eta \in (0, 1)$ for all t. For a global objective as per Eq. (2), if $p_i^t = p_i$ for all t,* *under FedAvg with exact local gradients*

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

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

The proof of Proposition 1 can be found in Appendix. It can be checked that if there exist $i, i' \in [m]$ such that $p_i \neq p_{i'}$, then $\lim_{t\to\infty} \boldsymbol{x}^t \neq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{u}_i \triangleq \boldsymbol{x}^*$; when $p_i = p$ for all $i \in [m]$, then $\lim_{t\to\infty} \boldsymbol{x}^t = \boldsymbol{x}^*$. In fact, the output of FedAvg may be far away from $\boldsymbol{x}^\star$ depending on p_i's and $\boldsymbol{u}_i$'s.

IV. ALGORITHM: FEDPBC

In this section, we propose FedPBC (*Federated Postponed Broadcast*, formally described in Algorithm 2) - a simple variant of FedAvg.

Algorithm 2: FedPBC

1 **Input:** T, $\boldsymbol{x}^0$, s, $\{\eta_t\}_{t=0,\cdots,T-1}$
2 The PS and each client initialize parameter $\boldsymbol{x}^0$;
3 **for** $t = 0, \cdots, T-1$ **do**
/* Let $\mathcal{A}^t$ denote all the clients with active communication links; */
4 **for** $i \in [m]$ **do**
5 Draw a fresh sample ξ_i^t;
6 $\boldsymbol{x}_i^{(t,0)} = \boldsymbol{x}_i^t$;
7 **for** $k = 0, \cdots, s-1$ **do**
8 $\boldsymbol{x}_i^{(t,k+1)} = \boldsymbol{x}_i^{(t,k)} - \eta_t \nabla \ell_i(\boldsymbol{x}_i^{(t,k)}; \xi_i^t)$;
9 **end**
10 $\boldsymbol{x}_i^{t++} = \boldsymbol{x}_i^{(t,s)}$;
11 Report $\boldsymbol{x}_i^{t++}$ to the PS;
12 **end**
/* On the PS. */
13 **if** $\mathcal{A}^t \neq \emptyset$ **then**
14 $\boldsymbol{x}^{t+1} \leftarrow \frac{1}{|\mathcal{A}^t|} \sum_{i\in\mathcal{A}^t} \boldsymbol{x}_i^{t++}$;
15 **else**
16 $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t$;
17 **end**
18 Multi-cast $\boldsymbol{x}^{t+1}$ to each client $i \in \mathcal{A}^t$;
19 **for** $m \in \mathcal{A}^t$ **do**
20 $x_i^{t+1} \leftarrow \boldsymbol{x}^{t+1}$;
21 **end**
22 **end**

The key difference of FedPBC from FedAvg is that we postpone the global model broadcasts to $\mathcal{A}^t$ till the end of each round. Postponing the global model broadcast introduces some staleness as the clients might 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 staleness is mild and there is no significant slowdown. Theoretical analysis and numerical results can be found in Sections V and VI, respectively.

Implicit gossiping among clients $\mathcal{A}^t$. From line 14 to line 22 of Algorithm 2, via the coordination of the PS, 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

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

The matrix is by definition *doubly-stochastic* and $W^{(t)} = I$ when $\mathcal{A}^t = \emptyset$ or $|\mathcal{A}^t| = 1$. We further note that this matrix can be *time-varying* even in expectation 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 different clients.

Let $M^{(t)} \triangleq \mathbb{E}\left[\left(W^{(t)}\right)^2\right]$ and $\mathbf{J} \triangleq \frac{1}{m}\mathbf{1}\mathbf{1}^\top$. Define as

$$\rho(t) \triangleq \lambda_2\left(M^{(t)}\right) \quad \text{and} \quad \rho \triangleq \max_t \rho(t). \tag{3}$$

Lemma 1 (Ergodicity). *Recall that $p_i^t \geq c$ for some constant $c \in (0,1)$. For each $t \geq 1$, it holds that $\rho \leq 1 - \frac{c^4[1-(1-c)^m]^2}{8}$.*

The following lemma will be used in the convergence analysis.

Lemma 2. *For any matrix $B \in \mathbb{R}^{d\times m}$, it holds that*

$$\mathbb{E}\left[\|B\left(\prod_{r=1}^{t} W^{(r)} - \mathbf{J}\right)\|_{\mathrm{F}}^2\right] \leq \rho^t \|B\|_F^2.$$

The proof of Lemma 2 follows the same outline as that in [12, Lemma 1].

Remark 1. *In Algorithm 2, each client does local computations even if its communication link is not active. Continuous local updates appear to be crucial. Numerical examples in Section VI show that bias persists when only the active clients do local computations. We leave as a future direction on how to remove the bias while maintaining local computation.*

V. CONVERGENCE RESULTS

A. Assumptions

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

Assumption 1 (Smoothness). *Each local gradient function $\nabla\ell_i(\theta)$ is L_i-Lipschitz, i.e.,*

$$\|\nabla\ell_i(\boldsymbol{x}_1) - \nabla\ell_i(\boldsymbol{x}_2)\|_2 \leq L_i \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2,$$

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

Assumption 2 (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.,

$$\mathbb{E}\left[\left\|\nabla \ell_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{x})\right\|_2^2\right] \leq \sigma^2,$$

where $\mathcal{F}^t$ denotes the sigma algebra generated by all the randomness up to iteration t.

Assumption 3. *There exists $F^* \in \mathbb{R}$ such that $F(\boldsymbol{x}) \geq F^*$ for all $\boldsymbol{x} \in \mathbb{R}^d$.*

Assumption 4 (Bounded Inter-client Heterogeneity)**.**

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

Assumptions, 1, 2 and 3 are standard in FL analysis [10], [13], [14]. Assumption 4 captures the heterogeneity across different users, and it is a more relaxed version (e.g., than [10], [15], [16].) Notably, different from [9], we do not assume fresh data per local update, and the unbiasedness in Assumption 2 is imposed for global rounds only.

B. Results

In this section, we formally state our key lemmas and main theorem. All proofs can be found in the full version [17].

Lemma 3 (Lemma 1 in [18])**.** *For $s \geq 1$, we have for all $\boldsymbol{x} \in \mathbb{R}^d$:*

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

where $\kappa \triangleq \max_i \frac{(1+\eta L_i)^s - 1 - s\eta L_i}{\binom{s}{2}(\eta L_i)^2}$.

For any $s \in \mathbb{N}$, κ is monotonic non-decreasing with respect to $\eta > 0$, where

$$\kappa \triangleq \frac{(1+\eta L)^s - 1 - s\eta L}{\binom{s}{2}(\eta L)^2}. \quad (4)$$

Remark 2. *Lemma 3 yields a simple upper bound on the perturbations incurred by multiple local steps. When $s = 1$, it holds that $\kappa = 0$. For $s \geq 2$, we always have $\kappa \geq 1$. Furthermore, due to the fact that κ is non-decreasing in η, it is true that $\kappa \leq \frac{e^c - 1 - c}{c^2/2}$ provided $\eta \leq \frac{c}{sL}$. In other words, we can treat κ as a constant as long as η is sufficiently small. Henceforth, we adopt the common convention that $\frac{\sqrt{2}}{\kappa s L} = \infty$ when $s = 1$. An immediate consequence of this convention is that, e.g., $\min\left\{\frac{1}{2s}, \frac{\sqrt{2}}{\kappa s L}, \frac{1-\sqrt{\rho}}{6\sqrt{2\rho}Ls^2}\right\} = \min\left\{\frac{1}{2s}, \frac{1-\sqrt{\rho}}{6\sqrt{2\rho}Ls^2}\right\}$.*

Let

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

Lemma 4 (Descent Lemma)**.** *Suppose Assumptions 1, 2, and 4 hold, under a choice of the learning rate $\eta \leq \frac{1}{2s}$, the following property holds for $t \geq 0$:*

$$\begin{aligned}
&\mathbb{E}\left[F(\bar{\boldsymbol{x}}^{t+1}) - F(\bar{\boldsymbol{x}}^t) \mid \mathcal{F}^t\right] \\
&\leq \sigma^2\eta^2 s^2\left[\kappa^2L^2 + 2L\left(\frac{1}{m} + \frac{\kappa^2L^2}{4}\right)\right] + 3\zeta^2\eta^2s^2\mathfrak{C} \\
&- \left\{\frac{s\eta}{4} - 3\eta^2s^2\left(\beta^2+1\right)\mathfrak{C}\right\}\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2 \\
&+ \left\{\eta s L^2 + 3\eta^2s^2L^2\mathfrak{C}\right\}\underbrace{\frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}^t\right\|_2^2}_{\text{consensus error}},
\end{aligned}$$

where $\mathfrak{C} \triangleq \kappa^2L^2 + 2L\left(1 + \frac{\kappa^2L^2}{4}\right)$.

Remark 3. *Lemma 4 can be proved via following the standard outline of SGD convergence analysis with non-convex functions and plugging in Lemma 3 to bound the perturbation arises from multiple local updates and non-fresh data per update. The consensus error term comes from Assumption 1 and enables us to connect our analysis of the aforementioned W matrix, where we borrow the insights from the analysis of gossiping algorithms. Formally, in matrix form, we use the following notions*

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

Equivalently, we can write down the consensus error in matrix form,

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

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

Lemma 5 (Consensus Error)**.** *Suppose the conditions in Lemma 4 are met, under a choice of the learning rate, $\eta \leq \min\left\{\frac{1}{2s}, \frac{\sqrt{2}}{\kappa s L}, \frac{1-\sqrt{\rho}}{6\sqrt{2\rho}Ls^2}\right\}$. The following property holds:*

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{X}^{(t)}\left(\mathbf{I} - \mathbf{J}\right)^2\|_{\mathrm{F}}\right] \\
&\leq 6\eta^2s^2\left[\frac{2\rho}{\left(1-\sqrt{\rho}\right)^2} + \frac{\rho}{1-\rho}\right]\sigma^2 + \frac{72\eta^2s^4\rho}{\left(1-\sqrt{\rho}\right)^2}\zeta^2 \\
&+ \frac{72\left(\beta^2+1\right)\eta^2s^4\rho}{\left(1-\sqrt{\rho}\right)^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right].
\end{aligned}$$

Now, we are ready to present our main theorem.

Theorem 1. *Suppose all the assumptions hold and $\mathfrak{C}$ as defined in Lemma 4, and choose a learning rate $\eta = c_0\sqrt{\frac{m}{sT}}$, where c_0 is a constant, for sufficiently large T such that*

$$\eta \le \min\left\{\frac{1}{24\left(\beta^2+1\right)\mathfrak{C}\left[1+\frac{18s^2L^2\rho}{\left(1-\sqrt{\rho}\right)^2}\right]+\frac{144(\beta^2+1)s^2L^2\rho}{\left(1-\sqrt{\rho}\right)^2}}, \frac{1}{2s}, \frac{\sqrt{2}}{\kappa sL}, \frac{1}{\rho s^3}, \frac{1-\sqrt{\rho}}{6\sqrt{2\rho}Ls^2}\right\},$$

the following property holds for Algorithm 2:

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \le O\left(\frac{8F(\bar{\boldsymbol{x}}^0)-8F^\star}{\sqrt{msT}}\right)$$
$$+\underbrace{O\left(16L\sqrt{\frac{s}{mT}}\sigma^2+8\sqrt{\frac{ms}{T}}\kappa^2L^2\left(1+\frac{L}{2}\right)\sigma^2\right)}_{\text{Stochastic gradient noise}}$$
$$+\underbrace{O\left(24\sqrt{\frac{ms}{T}}\left[\mathfrak{C}+\frac{24L^2}{\left(1-\sqrt{\rho}\right)^2}\right]\zeta^2+\frac{ms}{T}\frac{1728\mathfrak{C}L^2}{\left(1-\sqrt{\rho}\right)^2}\zeta^2\right)}_{\text{Client drift error}}$$
$$+\underbrace{O\left(\frac{ms}{T}\frac{144\rho L^2}{\left(1-\sqrt{\rho}\right)^2}\left(1+3\mathfrak{C}\right)\sigma^2\right)}_{\text{Intermittent participation error}}.$$

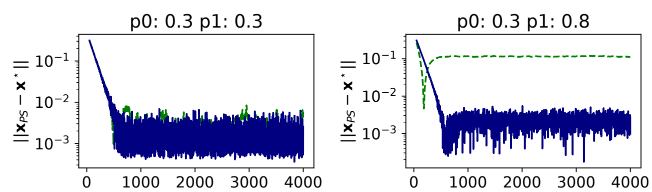
Remark 4. *Here, we remark on Theorem 1:*

1) **On the structures.** *Except for the first term, the remained terms can be grouped into three parts: the noise introduced by stochastic gradient, and the errors due to client drift (heterogeneity) and intermittent participation, each scaling with a different rate. To control the errors, we need a sufficiently small learning rate η that meets all the conditions mentioned above.*
2) **On stationary points of F.** *Theorem 1 says that $\bar{\boldsymbol{x}}^t$ in FedPBC converges to a stationary point of F asymptotically. In other words, the bias will be corrected towards the end. In contrast, we show in Proposition 1 that $\bar{\boldsymbol{x}}^t$ in FedAvg converges to a point that could be arbitrarily far away from the true optimum depending on p_i^t and data heterogeneity.*
3) **On the role of the activation lower bound c.** *It has been shown in Lemma 1 that $\rho \le 1-\frac{c^4[1-(1-c)^m]^2}{8}$. A greater c leads to a smaller ρ and thus a tighter bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F\left(\bar{\boldsymbol{x}}^t\right)\|_2\right]$. Note that FedPBC reduces to FedAvg with full-client participation, i.e., when $c=1$. In that case, our convergence rate becomes*
$$O\left(\frac{1}{\sqrt{msT}}+\sqrt{\frac{ms}{T}}+\frac{ms}{T}\right), \qquad (6)$$
which matches the FedAvg literature (see e.g., in [11]). We further note that because $\frac{\kappa}{s}$ can be treated as a constant, the order of convergence rate does not change.
4) **On linear speedup.** *It is trivial to see that the first two terms in Eq. (6) dominate when T is sufficiently large (e.g., $T \ge c_1m^3s^3$, where c_1 is some positive constant.) We shall see linear speedup w.r.t. the first term; however, the second term ultimately dominates all. Thus, it is unlikely that our algorithm achieves linear speedup, which is consistent with FedAvg literature, see e.g., in [3].*

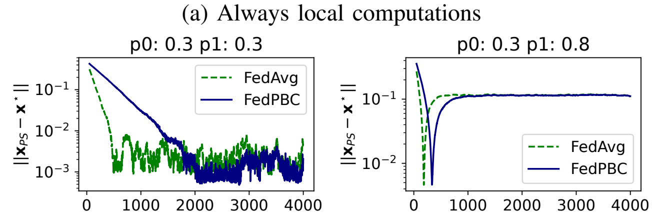
VI. Numerical Experiments

In this section, we present the numerical evaluations of the proposed algorithm and FedAvg. In each round, the PS will send an update request to each client. Client i will respond with probability p_i, which is unknown to both the PS and clients. This simulates unstable communications.

Counterexample. Here, we have $m = 100$ clients, each doing 30-steps local computations, communicating for 4000 rounds, and holding a local loss function $F_i(\boldsymbol{x}_i) = \frac{1}{2}\|\boldsymbol{x}_i-\boldsymbol{u}_i\|_2^2$, 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 $\eta = 0.0003$. In addition, we let the first 50 clients respond with probability p_0, whereas the second half with p_1 (to be specified later.)



(a) Always local computations

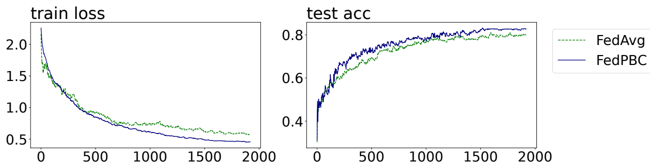


(b) Sampled local computations

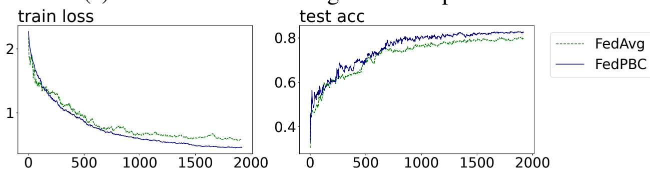
Fig. 2: Distance to the optimum $\|\boldsymbol{x}_{\text{PS}}-\boldsymbol{x}^\star\|_2$ in the counterexample in logarithmic scale.

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. 2, where $\boldsymbol{x}_{\text{PS}}^t \triangleq \boldsymbol{x}^t$ in Algorithm 2. As illustrated in Fig. 2a, FedPBC is unbiased and converges to the global optimum $\boldsymbol{x}^\star \triangleq \frac{1}{m}\sum_{i=1}^m \boldsymbol{u}_i$ in all the combinations of p_0 and p_1, matching our analysis, while FedAvg will instead converge to a different point observed from $\|\boldsymbol{x}_{\text{PS}}-\boldsymbol{x}^\star\|_2$ when $p_0 \ne p_1$. When $p_0 = p_1$, the two algorithms will converge to the same point consistent with our analysis. In sharp contrast, if we let only the sampled clients do local computations, the bias persists, which we leave as a future direction.

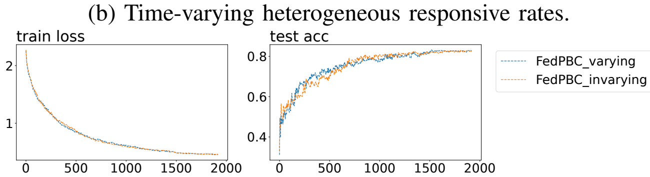
Synthetic $(1,1)$ data . In this simulation, we first follow [10] and construct Synthetic $(1,1)$ dataset as follows: we generate samples (X_i, Y_i) for each client i according to the



(a) Time-invariant heterogeneous responsive rates.



(b) Time-varying heterogeneous responsive rates.



(c) FedPBC evaluations under time-invariant and time-varying responsive rates.

Fig. 3: Synthetic $(1, 1)$ evaluations.

model $y = \arg\max\left(\text{softmax}\left(Wx + b\right)\right)$, where $x \in \mathbb{R}^{60}$, $W \in \mathbb{R}^{10\times 60}$, $b \in \mathbb{R}^{10}$. To characterize the non-i.i.d. data, we let $W_i \sim N\left(u_i, 1\right)$, $b_i \sim N\left(u_i, 1\right)$, $u_i \sim N\left(0, \alpha = 1\right)$, and $x_i \sim \mathcal{N}\left(v_i, \Sigma\right)$, where the covariance matrix is diagonal with $\sum_{j,j} = j^{-1.2}$. Each element in the mean vector v_i is drawn from $N\left(B_i, 1\right)$, where $B_i \sim \left(0, \beta = 1\right)$.

For the non-uniform link activation probabilities p_is, we consider two scenarios:

1) *Time-invariant* heterogeneous rates. Let $p_i^t = p_i^0 = 0.05$ for $1 \le i \le m/2$ and $p_j^t = p_j^0 = 0.9$ for $(m/2) + 1 \le j \le m$. In other words, we have two groups of clients, one responding with probability $p_i^0 = 0.05$, while the other one with probability $p_j^0 = 0.9$;
2) *Time-varying* heterogeneous rates. A uniformly distributed random variable, which is independent across clients and communication rounds, is imposed on each responsive rate per communication round. Formally, let $p_i^t = p_i^0 + X_i^t$ and $p_j^t = p_j^0 + X_j^t$, where $X_i^t, X_j^t \sim \mathcal{U}\left(-0.02, 0.02\right)$ for $1 \le i \le (m/2)$ and $(m/2) + 1 \le j \le m$. This ensures $c \triangleq \min_{t\in[T], i\in[m]} p_i^t = 0.03$.

The other auxiliary hyper-parameters are set as: client size $m = 30$, a constant learning rate η_0 tuned from $\{0.1, 0.5, 0.01, \ldots, 0.001, 0.005\}$, batch size: 100, local computation rounds: 25 for each $i \in [m]$, communication rounds: 1900.

Fig. 3a and Fig. 3b show that FedPBC consistently outperforms FedAvg. Moreover, Fig. 3c says that FedPBC converges to the same point in either setting, showing its ability to rectify the bias.

References

[1] 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.

[2] 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.

[3] 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

[4] 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.

[5] 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

[6] 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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5693–5700.

[16] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 9709–9758, 2021.

[17] M. Xiang, S. Ioannidis, E. Yeh, C. Joe-Wong, and L. Su, "Towards bias correction of fedavg over nonuniform and time-varying communications," *arXiv preprint arXiv:2306.00280*, 2023.

[18] L. Su, J. Xu, and P. Yang, "Federated learning in the presence of adversarial client unavailability," *arXiv preprint arXiv:2305.19971*, 2023.