PARTIAL PARTICIPATION: GUARANTEES OF LOCAL STEP CONVERGENCE FOR UNIFIED CLIENT SELECTION IN THE NON-CONVEX CASE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables model training across distributed devices without direct data sharing, making it a natural framework for large-scale learning systems. Its scalability is nevertheless constrained by two fundamental bottlenecks: server-side overload caused by the impossibility of involving all devices in each communication round, and channel overload caused by transmitting high-dimensional models. These challenges are traditionally addressed separately: the former through partial participation training paradigm, and the latter through compression of the transmitted information. In heterogeneous settings, however, standard partial participation schemes may lead to biased optimization dynamics and are difficult to analyze rigorously. While partial participation and compression have been extensively studied in isolation, their joint treatment under heterogeneity remains largely unresolved. In this work, we address this gap by developing a novel federated optimization algorithm that operates under partial participation with compressed communication. We establish convergence guarantees under general non-convex assumptions and empirically evaluate the proposed method on image classification benchmarks.

1 INTRODUCTION

In contemporary settings, machine learning is characterized by the availability of extremely large datasets and the growing computational cost of training, which often becomes prohibitively time-consuming. Distributed optimization addresses these limitations by splitting the computational workload across multiple devices that work in parallel (Verbraeken et al., 2020). Formally, the problem can be expressed as:

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{M}\sum_{m=1}^{M} f_m(x,\xi_m), \tag{1}$$

where M is the amount of devices, f_m is the local loss function on data $\xi_m \sim \mathcal{D}$, where $\mathcal{D}$ is the data distribution and x is the weights of the model.

However, in many real-world scenarios, data is generated and stored on local devices such as smartphones or wearable systems. Moreover, in certain settings, the data is privacy-sensitive and cannot be moved from the host device or local infrastructure due to legal and regulatory requirements (Shafik & Abubakari, 2026). To formalize this setting, we assume that each device $m \in \{1,\dots,M\}$ has access to data sampled from its own local distribution $\mathcal{D}_m$. Importantly, these distributions may differ across devices, and the corresponding local datasets can vary significantly in size. Federated learning (McMahan et al., 2017; Yang et al., 2019) addresses this challenge by keeping data on local devices and only transmitting model updates to a central server. The baseline solution is to aggregate local updates by computing their mean and then applying it to the model.

There is a question of the practical efficiency of such naive training will always be a pressing issue. In the federated learning setting, in addition to local computations, there is also a communication phase, which becomes a significant bottleneck in practice as it may be time-consuming or even unrealistic in some setups with inconsistent network connection. Several techniques exist to circumvent this issue.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Partial participation One fundamental issue is the impossibility of full communication. In large-scale deployments with thousands or millions of devices, the server cannot accept all clients responses at the same time. Partial participation addresses this issue by selecting only a random subset of clients to communicate in each round, while the other clients stay inactive (Li et al., 2020b). This approach reduces the server's load and naturally accommodates systems where devices are intermittently available.

This strategy is known to perform well under the assumption i.i.d. data across clients. However, in the presence of statistical heterogeneity, where local data distributions differ significantly, partial participation can lead to degraded performance due to the introduction of sampling bias. In particular, the subset of clients selected at each round may not be representative of the overall population, resulting in biased gradient estimates and slower or unstable convergence. Consequently, recent research has focused on designing more advanced client sampling strategies that explicitly account for data heterogeneity, aiming to mitigate this bias and improve both convergence speed and model generalization (Zhang et al., 2023).

Communication number. On the other hand, even when clients are reachable, transmitting updates between clients and the server is often slow and expensive, especially when models are large or network connectivity is limited (Kairouz et al., 2021). Federated Averaging (FedAvg) (McMahan et al., 2017), has become a foundational algorithm precisely because it tackles this issue. Its key feature is that each worker performs K local SGD steps between two consecutive communication rounds, in contrast to traditional distributed SGD where communication occurs after every step. This allows for a significant reduction in communication costs.

While frameworks such as FedAvg and SCAFFOLD (Karimireddy et al., 2021) may integrate both local steps and partial participation, they typically rely on uniform random sampling for client selection. This approach fails to account for device trust scores or selection biases, which leads to significant performance degradation in heterogeneous environments. Furthermore, although SCAFFOLD employs variance reduction techniques to improve theoretical convergence, these methods are often impractical or overly complex for real-world Federated Learning (FL) scenarios involving partial participation. Consequently, current variants lack the nuanced selective or weighted aggregation mechanisms required to handle the reliability and diversity of participating devices effectively.

This work aims to address this gap by developing a convergence analysis for local updates combined with partial participation considering heterogeneous environments.

2 Related Work

Local Steps. The integration of local steps, foundational to Local SGD (Stich, 2019), serves as a primary mechanism for mitigating communication bottlenecks by performing K local iterations before synchronization. Early investigations, such as the one done by Woodworth et al. (2020), provided the initial comparison between local and minibatch SGD. Subsequent theoretical advancements done by Gorbunov (2021) have attempted to establish a unified convergence analysis for these methods. However, these theoretical frameworks encounter significant difficulties under data heterogeneity. In heterogeneous settings, convergence typically requires highly restrictive assumptions, and the resulting performance estimates are often degraded due to the client drift, inherent in local steps

Within the specific framework of federated learning, the standard Federated Averaging (FedAvg) algorithm (McMahan et al., 2017) frequently suffers from poor convergence behavior when faced with non-IID data distributions across devices. To address the instability caused by local updates, more advanced algorithms like SCAFFOLD have been introduced. These frameworks incorporate variance reduction (VR)-based schemes to correct the local update directions, thereby improving convergence rates and providing more robust performance in real-world, heterogeneous federated environments.

Partial Participation and Client Weighting. In real-world federated systems, full participation is often impossible. This challenge has motivated research on different techniques to reduce communication (Alistarh et al., 2017; Stich, 2019). One of them is partial participation (Li et al., 2020b; Yang et al., 2021). Over the years, researchers have developed numerous strategies for selecting

which clients should participate (Rizk et al., 2020; Nguyen et al., 2020; Nishio & Yonetani, 2019; Lai et al., 2021).

Despite the proliferation of these strategies, each new sampling scheme requires its own convergence analysis, resulting in a fundamental trade-off between the generality of the setting and the strength of convergence guarantees. Early analyses, such as those of Li et al. (2020b), establish $O(1/K)$-type convergence rates under strongly convex and smooth objectives with unbiased client sampling; however, these results rely on restrictive assumptions, including bounded gradient variance and, in many cases, diminishing step sizes, without which a residual error persists. Subsequent works that relax the sampling assumptions, such as Jee Cho et al. (2022), explicitly introduce non-vanishing terms in their bounds to capture the effects of biased client selection and data heterogeneity. Although some approaches (Luo et al., 2021) report convergence without residual error terms, this is typically achieved by imposing stronger structural assumptions, such as bounded heterogeneity, or by incorporating additional mechanisms like variance reduction or gradient correction, thereby shifting the burden from the convergence result to the problem setup. More general frameworks, including unified analyses of arbitrary participation in non-convex settings (Wang & Ji, 2024), further demonstrate that convergence guarantees inherently involve non-vanishing terms that reflect participation irregularity and statistical heterogeneity. Taken together, these results indicate that either convergence bounds include non-vanishing error terms, or such terms can only be eliminated under restrictive assumptions, suggesting that residual error is an intrinsic feature of federated optimization with partial participation and heterogeneous data.

In contrast to these approaches, the proposed *PPBC* framework addresses the challenges of partial participation and client weighting from a different perspective (Bylinkin et al., 2025). This algorithm naturally supports a broad class of weighting and sampling strategies, enabling a unified treatment of heterogeneous participation schemes without requiring restrictive assumptions on client selection. Moreover, by design, *PPBC* exhibits robustness to periodic client disconnections, as delayed gradient contributions are preserved and incorporated in subsequent updates. As such, this algorithm provides a principled alternative to existing methods by mitigating the impact of partial participation while avoiding explicit non-vanishing correction terms in the optimization objective.

A natural extension is the incorporation of multiple local update steps between communication rounds, as commonly employed in federated optimization to reduce communication overhead. However, integrating local steps into the *PPBC* framework presents significant challenges.

3 Our Contribution

In this work, we bridge this gap by proposing a novel federated optimization algorithm *Partial Participation with Error Feedback and Local Steps (PP-EFLS)* that simultaneously addresses two primary bottlenecks: the impossibility of full client participation and the high overhead of frequent communication. The primary contributions of this paper are three-fold:

Algorithm Design. We develop a robust algorithm that integrates local update steps with a gradient-tracking mechanism, effectively mitigating client drift while accommodating inconsistent network availability.

Theoretical Analysis. We provide a rigorous convergence analysis of the proposed method, establishing its theoretical foundations under heterogeneous (non-IID) data distributions and non-convex objective.

Empirical Validation. We present extensive experimental results on image classification benchmarks comparing with existing baselines and demonstrating that our approach maintains stable convergence.

4 Preliminaries

In federated learning, the convergence analysis of algorithms traditionally relies on a set of basic assumptions regarding the properties of loss functions and data. First of all, the objective is assumed to be L-smooth. It is one of the key assumptions in optimization(Patel & Berahas, 2023).

Assumption 1 (L-smoothness)**.** *The function f is L-smooth on $\mathbb{R}^d$ with respect to the norm $\|\cdot\|$, i.e., it is differentiable and there exists $L > 0$ such that for any $x, y \in \mathbb{R}^d$ the following inequality holds:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Assumption 2 (Non-convexity)**.** *The function f is non-convex and bounded below, i.e., there exists $x^* \in \mathbb{R}^d$ such that*

$$f(x^*) = \inf_{x \in \mathbb{R}^d} f(x) > -\infty.$$

To capture data similarity, we introduce the following assumption:

Assumption 3 (Similarity)**.** *For some constants $\delta_1 \geq 0$ and $\delta_2 \geq 0$, clients satisfy*

$$\frac{1}{M}\sum_{m=1}^{M} \|\nabla f_m(x) - \nabla f(x)\|^2 \leq \delta_1 \|\nabla f(x)\|^2 + \delta_2.$$

This is a standard assumption in federated learning for quantifying data heterogeneity across clients and controlling client drift (Khaled et al., 2020; Karimireddy et al., 2021; Molodtsov et al., 2025). Next, we assume the following:

Assumption 4. *Each client m has access to an unbiased stochastic gradient satisfying $\mathbb{E}_{\xi_m}[\nabla f_m(x, \xi_m)] = \nabla f_m(x)$. Moreover, for all $x \in \mathbb{R}^d$, the variance is bounded by σ^2:*

$$\mathbb{E}_{\xi_m}\left[\|\nabla f_m(x, \xi_m) - \nabla f_m(x)\|^2\right] \leq \sigma^2.$$

Next, we consider that the weights $\{\pi_m\}_{m=1}^M$ lie on the regularized simplex. Namely,

$$\pi \in \Delta_M^1 \cap \left(\bigcap_{m=1}^{M} \left\{\pi : \mathbf{e}_m^\top \pi - \frac{\alpha}{M} \leq 0\right\}\right), \tag{2}$$

where $1 \leq \alpha \leq M$ is the regularization parameter and $\mathbf{e}_m$ denotes the m-th standard basis vector. This construction is widely used in different tasks(Mehta et al., 2025).

5 ALGORITHM AND THEORETICAL ANALYSIS

In this work, we address two fundamental challenges inherent to federated learning: handling partial client participation and incorporating multiple local update steps.

Partial participation is a paradigm where the server only uses gradients from clients that are active in the current round. When the server is forced to select only a subset of clients, this selection is almost always biased — some devices are chosen more frequently, others less often. This bias introduces a non-vanishing term in the convergence rate.

As discussed in the Section 2, the *PPBC* algorithm successfully addresses this issue. Its key idea is the introduction of a **gradient surrogate** — a variable that accumulates gradient information from all clients, including inactive ones, over an epoch. Thanks to this mechanism, even if a client does not participate in training for a while, its contribution is not lost but preserved in the surrogate and will be taken into account in the future.

Formally, *PPBC* maintains for each client m a variable g_m, which is updated according to the rule:

$$g_m \leftarrow g_m + (1 - \theta)\left(\frac{1}{M} - \hat{\pi}_m\right)\nabla f_m(x),$$

where $\hat{\pi}_m$ is the weight of the client, which can be zero for non-participating devices, and θ is a momentum parameter. The server stores the sum $g = \sum_m g_m$ and uses it to make updates. This technique helps *PPBC* to achieve convergence without non-vanishing terms even under biased sampling. In our algorithm, we employ a similar technique to handle partial participation.

Description. At the beginning of each epoch, the server computes the weights π^k according to the rule R (Line 5) — this implements partial participation. After that, the server initializes the surrogates $g_m^{k,0} = 0$ (Line 6) and sets the initial point of the epoch taking into account the accumulated surrogate from the previous epoch (Line 7):

$$x^{k,0} = x^{k-1,H^{k-1}} - \gamma g^{k-1,H^{k-1}}.$$

Algorithm 1 PP-EFLS

1: **Input:** Start point $x^{-1,H^{-1}} \in \mathbb{R}^d$, $g^{-1,H^{-1}} \in \mathbb{R}^d$, epochs number K, number of devices M
2: **Parameters:** Stepsize $\gamma > 0$, momentum $0 < \theta < 1$, regularization $1 \leq \alpha \leq M$, number of local steps $l \geq 1$, participation probabilities $q_m \in (0, 1]$
3: **for** $k = 0$ to $K - 1$ **do**
4: Initialize π^k
5: $\hat{\pi}^k = R^k(\pi^k)$
6: $g_m^{k,0} = 0$
7: $x^{k,0} = x^{k-1,H^{k-1}} - \gamma g^{k-1,H^{k-1}}$
8: Generate $H^k \sim \text{Geom}(p)$
9: **for** iterations $h = 0, \ldots, H^k - 1$ **do**
10: **for** devices $m = 1, \ldots, M$ in parallel **do**
11: $x_m^{k,h,0} = x^{k,h}$
12: **for** local steps $i = 0, \ldots, l - 1$ **do**
13: $\eta_m^{k,h,i} \sim \text{Be}(q_m)$
14: $x_m^{k,h,i+1} = x_m^{k,h,i} - \gamma \nabla f_m(x_m^{k,h,i}, \xi_m^{k,h,i})$
15: **end for**
16: $\nabla_m^{k,h} = \sum_{i=0}^{l-1} \frac{\eta_m^{k,h,i}}{q_m} \nabla f_m(x_m^{k,h,i}, \xi_m^{k,h,i})$
17: $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta)\left(\frac{1}{M} - \hat{\pi}_m^k\right) \nabla_m^{k,h}$
18: **end for**
19: **for** each m where $\hat{\pi}_m^k \neq 0$ **do**
20: Send $\nabla_m^{k,h}$ to the server
21: **end for**
22: $x^{k,h+1} = x^{k,h} - \gamma \left((1 - \theta) \sum_{m=1}^M \hat{\pi}_m^k \nabla_m^{k,h} + \theta g^{k-1,H^{k-1}}\right)$
23: **end for**
24: **for** each m **do**
25: g_m^{k,H^k} to the server
26: **end for**
27: $g^{k,H^k} = \sum_{m=1}^M g_m^{k,H^k}$
28: **end for**
29: **return** $x^{K-1,H^{K-1}}$

When each client m performs l local steps (Lines 12-15). While adaptive selection of local steps l is possible, we use a constant value. This choice is optimal for non-convex problems Qin et al. (2023). With probability q_m, the client participates in the computations (Line 13); this is modeled by the indicator variable η_m^{khi}. This parameter captures the practical scenario where devices may periodically become inactive. These clients not only stop sending updates, but also stop local processing. To account for the fact that clients participate with different probabilities q_m, we normalize the computed gradients by these factors. This ensures that the expected value of each corrected gradient $\nabla_m^{k,h}$ equals the true gradient $\nabla f_m(x_m^{k,h})$, i.e., $\mathbb{E}_{\eta_m^{k,h,i}}\left[\frac{\eta_m^{k,h,i}}{q_m} \nabla f_m(x_m^{k,h,i}, \xi_m^{k,h,i})\right] = \nabla f_m(x_m^{k,h,i})$.

After completing l local steps, the client sends the sum of the corrected gradients ∇_m^{kh} to the server (Line 16), then updates the global model $x^{k,h}$ (Line 22).

Surrogate accumulation. A natural question arises when incorporating local steps into the surrogate-based framework: how should the server aggregate the local gradient information from each client? In Algorithm 1, we define

$$\nabla_m^{k,h} = \sum_{i=0}^{l-1} \frac{\eta_m^{k,h,i}}{q_m} \nabla f_m(x_m^{k,h,i}, \xi_m^{k,h,i}),$$

i.e., the sum over all l local stochastic gradients.

A naive approach would be to use the averaged surrogate or, for instance, only the last gradient computed over the local epoch. However, from a practical perspective, this would mean that the algorithm performs only a single effective local step, thereby failing to realize the full benefits of

the local steps technique. This motivates our choice of summing the gradient over the entire local epoch, as opposed to any alternative aggregation strategy.

Analysis. We introduce virtual sequences to derive convergence rates of the algorithm:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^{M} g_m^{k,h}.$$

From the update rules of *PP-EFLS*, the virtual sequence satisfies the following recurrence:

$$\tilde{x}^{k,h+1} = \tilde{x}^{k,h} - \gamma \left[(1-\theta)\frac{1}{M} \sum_{m=1}^{M} \nabla_m^{k,h} + \theta g^{k-1,H^{k-1}} \right].$$

Subsequently, we will conduct the analysis on this virtual sequence. This choice is motivated by the fact that the update accounts for the accumulated surrogates from the previous epoch, while now incorporating all devices rather than only those participating in the aggregation.

One of the challenges in analyzing convergence with local steps is understanding how the gradient surrogate behaves in this case. A natural concern is whether the surrogate becomes dependent on the number of local steps, potentially accumulating an error that grows uncontrollably with l. To address this challenge, we establish the following lemma.

Lemma 1. *Suppose Assumptions 1,3, and 4 hold. If* $\gamma \leq \frac{1}{3\cdot\sqrt{33(\delta_1+1)}\cdot lL}$*, then for Algorithm 1 we have:*

$$\begin{aligned}\mathbb{E}_{H^k}\mathbb{E}_{\eta_m^{k,0}}\mathbb{E}_{\xi_m^{k,0}}\ldots\mathbb{E}_{\eta_m^{k,h}}\mathbb{E}_{\xi_m^{k,h}}\|g^{k,H^k}\|^2 \leq & 24(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2 q_{min}}\mathbb{E}_{H^k}\|\nabla f(x^{k,H^k})\|^2 \\ & + 48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2 q_{min}} + 24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2 q_{min}}.\end{aligned}$$

The non-stochastic terms exhibit a quadratic dependence on l, which is natural given the accumulation of the surrogate gradient across local steps. However, owing to the independence of the random variables $\xi_m^{k,h,i}$ and $\eta_m^{k,h,i}$, the variance term scales linearly with l rather than l^2. For more details, see Appendix, Lemma 7.

Another problem is the analysis of the following term:

$$\left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i}) \right\rangle.$$

When local steps are disabled, we have $x_m^{k,h,i} = x^{k,h}$, and taking expectations gives the familiar term $-\langle\nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h})\rangle$, which is a crucial term in the analysis. In contrast, enabling local steps introduces an extra deviation term

$$\left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i}) - \nabla f_m(x_m^{k,h}) \right\rangle.$$

The deviation inherently involves the impact of heterogeneity of the local functions. In our analyzes we demonstrate that the deviation can be bounded effectively. This leads to the following theorem.

Theorem 1. *Suppose Assumptions 1, 2, 3, and 4 hold. Then for Algorithm 1 with parameters satisfying*

$$\gamma \leq \frac{p\cdot q_{min}}{33Kl\alpha(\delta_1+1)}, \qquad \theta \leq \frac{\gamma L p^2}{2}$$

we have:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 \leq & \frac{16}{\gamma K l}\mathbb{E}\left[f(x^{0,0}) - f(x^*)\right] \\ & + 384L\gamma\frac{\alpha\delta_2}{pq_{min}} + 24L\gamma\frac{l\delta_2}{pq_{min}} + 192L^2\gamma^2\frac{\alpha\delta_2 l^2}{p^3 q_{min}} \\ & + 192L\gamma\frac{\alpha\sigma^2}{lpq_{min}} + 16L\gamma\frac{\sigma^2}{pq_{min}} + 96L^2\gamma^2\frac{l\alpha\sigma^2}{p^3 q_{min}}.\end{aligned}$$

Here $q_{\min} = \min\limits_{m=1,\ldots,M} q_m$. To obtain this bound, it was necessary to estimate the term with $\|g^{k,h}\|^2$. The use of local steps reduces the coefficient of this term by a factor of l. This allows us to eliminate the dependence on l in the terms $L\gamma\frac{\alpha\delta_2}{pq_{\min}}$ and $L\gamma\frac{\alpha\sigma^2}{lpq_{\min}}$. For more details, see Appendix, Theorem 2. This theorem immediately implies the following:

Corollary 1. *Under the conditions of Theorem 1, Algorithm 1 needs*

$$\mathcal{O}\left(\frac{\Delta L\alpha\delta_1}{pq_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lpq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{pq_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2pq_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lpq_{min}\varepsilon^4}\right) \text{ iterations and}$$

$$\mathcal{O}\left(M\left(\frac{M}{\min\limits_k C^k}\right)^2 \cdot \left(\frac{\Delta L\alpha\delta_1}{q_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{q_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2q_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lq_{min}\varepsilon^4}\right)\right)$$

number of devices communications to achieve ε-accuracy. Here $\Delta = f(x^{0,0}) - f(x^*)$, $\varepsilon^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2$, C^k is the number of devices participating in the k-th iteration.

A natural question is whether the obtained bounds are tight. First, we note that when no local steps are performed, the bound reduces to the one previously established in Bylinkin et al. (2025): $O\left(M\left(\frac{M}{\min_{k,h} C_{k,h}}\right)^2 \frac{1}{\min_{1\le m\le M} q_m}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{\varepsilon^4}\right)\right)$. Notably, the use of local steps leads to an interesting phenomenon. If the parameter α is not excessively large, its influence can be compensated by increasing the number of local steps l. As a result, the dominant terms associated with δ_1 and δ_2 no longer depend on α: $\mathcal{O}\left(\frac{\sigma^2}{l\varepsilon^4}\right)$ and $\mathcal{O}\left(\frac{\delta_2}{\varepsilon^4}\right)$. Finally, regarding algorithms with local steps, our estimate is of the same order as the results obtained by other authors (Khaled et al., 2020; Karimireddy et al., 2021). For example, Karimireddy et al. (2021) provides the following communication complexity bound for the non-convex setting: $\mathcal{O}\left(\frac{\delta_1}{\varepsilon^4} + \frac{\delta_2}{\varepsilon^4} + \frac{\sigma^2}{l\varepsilon^4}\right)$, which is consistent with our result.

6 Experiments

6.1 Setup

To validate the theoretical findings in practice, we conduct experiments.

Task. We focus on the image classification task using the *CIFAR-10 dataset* (Krizhevsky et al., 2009). For the model, we adopt *ResNet-18* (He et al., 2015), which serves as a standard benchmark architecture for optimization problems.

We consider the following federated learning setup. The training data is distributed across M = 10 clients under different data distributions, generated via dirichlet distribution with varying α parameters:

- $\alpha = 0.1$: This distribution can be considered the one with the most heterogenious data;
- $\alpha = 5$: This distribution can be considered the one with moderately heterogenious data;
- $\alpha = 10000$: This distribution can be considered the one with the least heterogenious data.

Baseline choice. We evaluate our proposed method against *FedAvg* and *FedProx* as representative baselines in federated optimization (McMahan et al., 2017; Li et al., 2020a). *FedAvg* serves as the canonical reference algorithm in federated learning, establishing a standard paradigm based on local stochastic optimization followed by server-side model averaging. In contrast, *FedProx* extends *FedAvg* by incorporating a proximal regularization term that explicitly addresses client heterogeneity and stabilizes local updates under non-IID data distributions. Comparing against both methods therefore enables a comprehensive evaluation of whether the proposed algorithm provides improvements over the foundational approach as well as over established techniques specifically designed to mitigate data heterogeneity.

Local steps. As this paper focuses on convergence with respect to local steps, we compare the algorithm's convergence across different numbers of local steps: 1, 5, and 10. This provides an accurate representation of the overall structure while avoiding the negative effects of overfitting.

Client sampling strategies. We adopt Power-of-Choice (*PoC*) (Cho et al., 2020) and Federated Class-Balanced Sampling (*FedCBS*) (Zhang et al., 2023) as client sampling strategies due to their complementary strengths in addressing the efficiency–accuracy trade-off inherent in federated learning. *PoC* introduces an informed selection mechanism in which a subset of clients is first sampled randomly, followed by the selection of those with the most favorable local loss or gradient characteristics. This strategy reduces variance in client updates and can accelerate convergence by prioritizing more informative participants. In contrast, *FedCBS* explicitly accounts for client heterogeneity by organizing or weighting client participation to improve representativeness across the population.

Evaluation Protocol. We compare the baseline algorithms in terms of accuracy under the same uplink budget. The number of uplink communications matters because uplink is typically the tightest bottleneck in federated learning: client devices usually sit on slower and less reliable connections – making the uplink count a direct proxy for the real-world cost of deploying the algorithm at scale.

Communication details. For the first experiment with partial participation, we employed a strategy where only 3 clients are available for iterations during each epoch, and from these we select one client to communicate with per iteration, for a total of 3 iterations per epoch. This method was implemented for the 3 algorithms to compare them under the same client sampling strategy.

The next experiment compares partial participation results with full participation results. It is important to note that *PP-EFLS* requires additional uplink communications compared to *FedAvg* and *FedProx* under full participation: while the former requires 3 uplink communications during the 3 iterations described above and 10 uplink communications at the end of the epoch, the former require only 10 uplink communications, equal to the amount of clients. To facilitate a fair evaluation, we adjusted the number of epochs so that the algorithms have the same uplink communication budgets.

For more details and plots you may look into Section D of Appendix.

6.2 Experiment results under partial participation

In the tables below, we compare the best accuracy over 100 epochs of the algorithms, averaged over several runs, across different numbers of local steps (in the tables below LS1, LS5, LS10 stand for 1, 5, 10 local steps respectively) and different Dirichlet distribution settings for the datasets, employing the partial participation structure and client sampling strategies described in the previous subsection.

Algorithm	LS1	LS5	LS10
PP-EFLS	0.27	0.45	0.64
FedAvg	0.19	0.25	0.4
FedProx	0.23	0.42	0.5

Table 1: PoC; $\alpha = 0.1$

Algorithm	LS1	LS5	LS10
PP-EFLS	0.33	0.52	0.66
FedAvg	0.28	0.45	0.56
FedProx	0.25	0.45	0.54

Table 2: FedCBS; $\alpha = 0.1$

Algorithm	LS1	LS5	LS10
PP-EFLS	0.45	0.67	0.80
FedAvg	0.35	0.53	0.66
FedProx	0.37	0.53	0.64

Table 3: PoC; $\alpha = 5$

Algorithm	LS1	LS5	LS10
PP-EFLS	0.45	0.69	0.81
FedAvg	0.37	0.57	0.68
FedProx	0.36	0.57	0.68

Table 4: FedCBS; $\alpha = 5$

Algorithm	LS1	LS5	LS10
PP-EFLS	0.45	0.70	0.80
FedAvg	0.37	0.53	0.64
FedProx	0.37	0.54	0.64

Table 5: PoC; $\alpha = 10000$

Algorithm	LS1	LS5	LS10
PP-EFLS	0.47	0.68	0.81
FedAvg	0.38	0.55	0.67
FedProx	0.38	0.55	0.65

Table 6: FedCBS; $\alpha = 10000$

Discussion. From the tables above, we see that *PP-EFLS* consistently outperforms the baselines across all configurations. The advantage is notably larger under higher heterogeneity ($\alpha = 0.1$), and while the gap narrows with more homogeneous data, it never fully closes.

All three algorithms improve monotonically with more local steps, since more local computation per round generally accelerates convergence within a fixed epoch budget. *PP-EFLS* shows the steepest improvement curve, whereas the baselines show more irregular growth — for instance, *FedAvg* barely improves from LS1 to LS5 under $\alpha = 0.1$ before jumping sharply at LS10.

Both *PoC* and *FedCBS* sampling strategies yield similar overall results; however, *FedCBS* tends to perform somewhat better in the more challenging regime of higher heterogeneity and fewer local steps.

6.3 Comparison of partial participation PP-EFLS with full participation baselines

In this section, we compare the results obtained under full participation with the results of *PP-EFLS* under partial participation, which were presented in the previous section. Understanding how far the behavior of a method under partial participation departs from its full-participation counterpart is essential for assessing its practical relevance.

			$\alpha = 0.1$			$\alpha = 5$			$\alpha = 10\,000$		
Algorithm	Part-n	# Up.	LS1	LS5	LS10	LS1	LS5	LS10	LS1	LS5	LS10
PP-EFLS	Part.	130	0.33	0.52	0.66	0.45	0.69	0.81	0.47	0.70	0.81
FedAvg	Full	130	0.24	0.50	0.71	0.5	0.77	0.82	0.5	0.79	0.83
FedProx	Full	130	0.36	0.61	0.69	0.51	0.77	0.82	0.51	0.79	0.83

Table 7: Experiment results for algorithms under full participation

In the table above Part-n. stands for participation and indicates, whether participation was full or partial; # Up. stands for the number of uplink communications performed by the algorithm, computed over the number of epochs each algorithm was run for (100 for *PP-EFLS* and 130 for *FedAvg* and *FedProx*).

Discussion. Overall, the comparison results follow an expected pattern. Under higher heterogeneity and fewer local steps, *FedAvg* achieves lower accuracy than both *FedProx* with full participation and *PP-EFLS* with partial participation, for the same number of communications. At higher values of α, however, with lower heterogeneity and more local steps, *FedAvg*'s accuracy becomes comparable to both *FedProx* with full participation and *PP-EFLS* with partial participation. *PP-EFLS* under partial participation does not surpass *FedProx*'s accuracy under full participation, even though it substantially outperforms *FedProx* under partial participation. This may be a consequence of the strict partial-participation regime, in which only up to 30% of clients are sampled per round. Under a less restrictive regime, it is possible that *PP-EFLS* would outperform *FedProx* in terms of accuracy.

7 Conclusion

In this paper, we proposed *PP-EFLS*, a federated optimization algorithm that simultaneously addresses biased partial participation under heterogeneous data and high communication costs. By integrating local update steps with a gradient surrogate mechanism, our method preserves unbiased convergence properties while substantially reducing communication overhead. We provided a unified convergence analysis under general non-convex assumptions, demonstrating that the algorithm achieves provable guarantees without introducing non-vanishing bias terms. We further validated these guarantees empirically, comparing *PP-EFLS* against *FedAvg* and *FedProx* across varying local steps, heterogeneity levels, and sampling strategies; *PP-EFLS* consistently achieved the highest accuracy in all configurations, with the largest gains under severe heterogeneity ($\alpha = 0.1$). Moreover, *PP-EFLS* under partial participation either outperforms or matches *FedAvg* under full participation, whilst having the same uplink budget, with heterogeneous setup. Overall the experiments confirm that its theoretical benefits translate into practical performance improvements.

AI USE STATEMENT

A large language model was used solely for language refinement and copy-editing of the manuscript text – e.g., improving grammar, clarity, and sentence structure. It was not used for research ideation, experimental design, data analysis, or generation of scientific content or claims. All technical content, results, and conclusions are the authors' own, and the authors take full responsibility for the accuracy and integrity of the paper.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding, 2017. URL https://arxiv.org/abs/1610.02132.

Zeyuan Allen-Zhu. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization, 2018. URL https://arxiv.org/abs/1802.03866.

Dmitry Bylinkin, Daniil Medyakov, Vladimir Aletov, Nail Bashirov, Michael Diskin, and Aleksandr Beznosikov. A robust training method for federated learning with partial participation. 2025.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies, 2020. URL https://arxiv.org/abs/2010.01243.

Eduard Gorbunov. Distributed and stochastic optimization methods with gradient compression and local steps, 2021. URL https://arxiv.org/abs/2112.10645.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10351–10375. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/jee-cho22a.html.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2021. URL https://arxiv.org/abs/1912.04977.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning, 2021. URL https://arxiv.org/abs/1910.06378.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data, 2020. URL https://arxiv.org/abs/1909.04715.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*, pp. 19–35, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020a. URL https://arxiv.org/abs/1812.06127.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020b. URL https://arxiv.org/abs/1907.02189.

Bing Luo, Wenli Xiao, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling, 2021. URL https://arxiv.org/abs/2112.11256.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Ronak Mehta, Jelena Diakonikolas, and Zaid Harchaoui. Drago: Primal-dual coupled variance reduction for faster distributionally robust optimization, 2025. URL https://arxiv.org/abs/2403.10763.

Gleb Molodtsov, Daniil Medyakov, Sergey Skorik, Nikolas Khachaturov, Shahane Tigranyan, Vladimir Aletov, Aram Avetisyan, Martin Takáč, and Aleksandr Beznosikov. Bant: Byzantine antidote via trial function and trust scores, 2025. URL https://arxiv.org/abs/2505.07614.

Hung T. Nguyen, Vikash Sehwag, Seyyedali Hosseinalipour, Christopher G. Brinton, Mung Chiang, and H. Vincent Poor. Fast-convergent federated learning, 2020. URL https://arxiv.org/abs/2007.13137.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–7. IEEE, May 2019. doi: 10.1109/icc.2019.8761315. URL http://dx.doi.org/10.1109/ICC.2019.8761315.

Vivak Patel and Albert S. Berahas. Gradient descent in the absence of global lipschitz continuity of the gradients, 2023. URL https://arxiv.org/abs/2210.02418.

Tiancheng Qin, S. Rasoul Etesami, and César A. Uribe. The role of local steps in local sgd. *Optimization Methods and Software*, 40(5):1017–1043, August 2023. ISSN 1029-4937. doi: 10.1080/10556788.2023.2241151. URL http://dx.doi.org/10.1080/10556788.2023.2241151.

Elsa Rizk, Stefan Vlaski, and Ali H. Sayed. Optimal importance sampling for federated learning, 2020. URL https://arxiv.org/abs/2010.13600.

Wasswa Shafik and Mussa Abubakari. *Regulatory Frameworks: HIPAA, GDPR, and Compliance in Federated Learning*, pp. 99–123. 02 2026. ISBN 978-3-032-03984-2. doi: 10.1007/978-3-032-03985-9_4.

Sebastian U. Stich. Local sgd converges fast and communicates little, 2019. URL https://arxiv.org/abs/1805.09767.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. A survey on distributed machine learning. *ACM Computing Surveys*, 53(2):1–33, March 2020. ISSN 1557-7341. doi: 10.1145/3377454. URL http://dx.doi.org/10.1145/3377454.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation, 2024. URL https://arxiv.org/abs/2205.13648.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local sgd for heterogeneous distributed learning, 2020. URL https://arxiv.org/abs/2006.04735.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning, 2021. URL https://arxiv.org/abs/2101.11203.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications, 2019. URL https://arxiv.org/abs/1902.04885.

Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, and Hai Li. Fed-cbs: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction, 2023. URL https://arxiv.org/abs/2209.15245.

A AUXILIARY NOTATION AND FACTS

In this section, we introduce additional notation and auxiliary facts to simplify the proof. Fix iteration k, epoch h and define the following quantities:

- $\xi_m^{k,h,i}$ - noise at the i-th local step over all local steps,
- $\eta_m^{k,h,i}$ - indicator whether computations occur on device m, over all local steps.

Table 8: Notation used in the proofs.

Symbol	Definition	Description
$\mathbb{E}_{\xi_m^{k,h}}$	$\mathbb{E}_{\xi_m^{k,h,0}} \dots \mathbb{E}_{\xi_m^{k,h,l-1}}$	Expectation over ξ for all devices m and local steps
$\mathbb{E}_{\eta_m^{k,h}}$	$\mathbb{E}_{\eta_m^{k,h,0}} \dots \mathbb{E}_{\eta_m^{k,h,l-1}}$	Expectation over η for all devices m and local steps
$q_{\min}$	$\min\limits_{m=1,\dots,M} q_m$	$\mathbb{E}_{\eta_m^{k,h}}\left[\eta_m^{k,h,i}\right] = q_i$, i.e., $\eta_m^{k,h,i} \sim \mathrm{Be}(q_i)$ by definition

Next, we are ready to present the following auxiliary results. We begin by stating two classical inequalities.

Proposition 1 (Young's and Cauchy-Schwarz inequalities)**.** *Suppose $x, y, \{a_i\}_{i=1}^n \in \mathbb{R}^d$, $c > 0$. Then,*

$$\|x+y\|^2 \le (1+c)\|x\|^2 + \left(1+\frac{1}{c}\right)\|y\|^2 \quad \textit{or in other form} \quad \langle x, y\rangle \le \frac{c}{2}\|x\|^2 + \frac{1}{2c}\|y\|^2,$$

$$\left\|\sum_{i=1}^n a_i\right\|^2 \le n\sum_{i=1}^n \|a_i\|^2.$$

Additionally, we will need several lemmas concerning the properties of the geometric distribution. These two will be used in the proof of Lemma 7:

Lemma 2. *Let H^k be a geometrically distributed random variable with parameter $p \in (0,1]$ and $x^{k,i}$ - points produced by Algorithm 1. Then the following inequality holds:*

$$\sum_{j\ge 0} p(1-p)^j \sum_{i=0}^{j-1} (1+p/2)^{j-i-1} \left\|\nabla f(x^{k,i})\right\|^2 \le \frac{2}{p}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2.$$

Proof. Expanding the left-hand side, we obtain

$$\begin{aligned}
&\sum_{j\ge 0} p(1-p)^j \sum_{i=0}^{j-1} (1+p/2)^{j-i-1} \left\|\nabla f(x^{k,i})\right\|^2 \\
&= p\left[(1-p)\left\|\nabla f(x^{k,0})\right\|^2 + (1-p)^2\left\{(1+p/2)\left\|\nabla f(x^{k,0})\right\|^2 + \left\|\nabla f(x^{k,1})\right\|^2\right\} + \cdots\right] \\
&= p(1-p)\left[(1-p)^0(1+p/2)^0 + (1-p)(1+p/2) + \cdots\right]\left\|\nabla f(x^{k,0})\right\|^2 \\
&\quad + p(1-p)^2\left[(1-p)^0(1+p/2)^0 + (1-p)(1+p/2) + \cdots\right]\left\|\nabla f(x^{k,1})\right\|^2 + \cdots \\
&\le \sum_{l\ge 0}(1-p)^l\left(1+\frac{p}{2}\right)^l \sum_{j\ge 0} p(1-p)^{j+1}\left\|\nabla f(x^{k,j})\right\|^2 \\
&\le \frac{1}{1-(1-p)\left(1+\frac{p}{2}\right)}\sum_{j\ge 0} p(1-p)^j\left\|\nabla f(x^{k,j})\right\|^2 = \frac{2}{p(p+1)}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
&\le \frac{2}{p}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2,
\end{aligned}$$

where the last inequality follows from $\frac{2}{p(p+1)} \le \frac{2}{p}$ for $p > 0$. This completes the proof. □

Lemma 3. *Let H^k be a geometrically distributed random variable with parameter $p \in (0, 1]$. Then the following inequality holds:*

$$\sum_{j\geq 0} p(1-p)^j \sum_{i=0}^{j-1} \left(1+\frac{p}{2}\right)^{j-i-1} \leq \frac{4}{p}.$$

Proof.

$$\begin{aligned}\sum_{j\geq 0} p(1-p)^j \sum_{i=0}^{j-1} \left(1+\frac{p}{2}\right)^{j-i-1} &\leq p\sum_{j\geq 0}(1-p)^j \cdot j \cdot \left(1+\frac{p}{2}\right)^j \leq p\sum_{j\geq 0} j\left(1-\frac{p}{2}\right)^j \\ &= p\cdot \frac{1-\frac{p}{2}}{\left(1-\left(1-\frac{p}{2}\right)\right)} = \frac{4(1-\frac{p}{2})}{p} \leq \frac{4}{p}.\end{aligned}$$

This completes the proof. □

The final lemma is taken from Allen-Zhu (2018). This property is necessary in Theorem 1, as it allows us to bound the expectations that emerge when performing a single descent step.

Lemma 4 (Allen-Zhu (2018))**.** *Given a sequence $D_0, D_1, \ldots, D_N \in \mathbb{R}$, where $N \sim \mathrm{Geom}(p)$. Then the following inequality holds:*

$$\mathbb{E}_N[D_{N-1}] = pD_0 + (1-p)\,\mathbb{E}_N[D_N].$$

B LEMMAS

This section is dedicated to the technical lemmas that bound the quantities appearing in the proof of the main theorem (Theorem 1). The first of the three lemmas estimates the distance each client moves away from the initial point which is common to all of them at the beginning of a local epoch.

Lemma 5. *Suppose Assumptions 1, 3, and 4 hold. If $\gamma \leq \frac{1}{\sqrt{3}\cdot lL}$, then for Algorithm 1 we have:*

$$\mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|x_m^{k,h,i}-x^{k,h}\right\|^2\right] \leq 33\gamma^2 l^2(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2 + 33\gamma^2 l^2\delta_2 + 11\gamma^2 l\sigma^2.$$

Proof. Let us fix h and k and define $\Delta_m^i := x_m^{k,h,i} - x^{k,h}$. Starting with the update rule from Line 14 of Algorithm 1, we have

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 &= \mathbb{E}_{\xi_m^{k,h}}\left\|x_m^{k,h,i-1} - \gamma\nabla f_m(x_m^{k,h,i-1},\xi_m^{k,h,i-1}) - x^{k,h}\right\|^2 \\ &= \mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1} - \gamma\nabla f_m(x_m^{k,h,i-1},\xi_m^{k,h,i-1})\right\|^2.\end{aligned}$$

Next, we decompose the stochastic gradient into its deterministic part and the zero-mean noise:

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 \\ &= \mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1} - \gamma\nabla f_m(x_m^{k,h,i-1}) + [\gamma\nabla f_m(x_m^{k,h,i-1}) - \gamma\nabla f_m(x_m^{k,h,i-1},\xi_m^{k,h,i-1})]\right\|^2 \\ &\leq \mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1} - \gamma\nabla f_m(x_m^{k,h,i-1})\right\|^2 + \gamma^2\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i-1}) - \nabla f_m(x_m^{k,h,i-1},\xi_m^{k,h,i-1})\right\|^2 \\ &\quad + 2\gamma\mathbb{E}_{\xi_m^{k,h}}\left\langle\Delta_m^{i-1} - \gamma\nabla f_m(x_m^{k,h,i-1}), \nabla f_m(x_m^{k,h,i-1}) - \nabla f_m(x_m^{k,h,i-1},\xi_m^{k,h,i-1})\right\rangle \\ &\leq \mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1} - \gamma\nabla f_m(x_m^{k,h,i-1})\right\|^2 + \gamma^2\sigma^2,\end{aligned}$$

where the last inequality is a direct consequence of Assumption 4. Applying Young's inequality with parameter c and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 \leq \mathbb{E}_{\xi_m^{k,h}}&\left(1+\frac{1}{c}\right)\left\|\Delta_m^{i-1}\right\|^2 + \gamma^2\sigma^2 \\ &+ 3\gamma^2(1+c)\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i-1}) - \nabla f_m(x^{k,h})\right\|^2 \\ &+ 3\gamma^2(1+c)\left\|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\right\|^2 \\ &+ 3\gamma^2(1+c)\left\|\nabla f(x^{k,h})\right\|^2.\end{aligned}$$

By averaging across all devices and using Assumption 1, we get

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 \leq & \left(1+\frac{1}{c}\right)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1}\right\|^2+\gamma^2\sigma^2\\ &+3\gamma^2(1+c)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i-1})-\nabla f_m(x^{k,h})\right\|^2\\ &+3\gamma^2(1+c)\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m(x^{k,h})-\nabla f(x^{k,h})\right\|^2\\ &+3\gamma^2(1+c)\left\|\nabla f(x^{k,h})\right\|^2\\ \leq & \left(1+\frac{1}{c}+3\gamma^2L^2(1+c)\right)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1}\right\|^2+\gamma^2\sigma^2 \qquad (3)\\ &+3\gamma^2(1+c)\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m(x^{k,h})-\nabla f(x^{k,h})\right\|^2 \qquad (4)\\ &+3\gamma^2(1+c)\left\|\nabla f(x^{k,h})\right\|^2.\end{aligned}$$

Now we need to choose γ and c such that the term in 3 might be estimated as follows: $\left(1+\frac{1}{c}+3\gamma^2L^2(1+c)\right)\leq\left(1+\frac{2}{l-1}\right)$. To this end, we set $c=l-1$ and $\gamma\leq\frac{1}{\sqrt{3}\cdot lL}$. Additionally, we can apply Assumption 3 to the term 4 to get:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 \leq & \left(1+\frac{2}{l-1}\right)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^{i-1}\right\|^2+\gamma^2\sigma^2\\ &+3\gamma^2l(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+3\gamma^2l\delta_2.\end{aligned}$$

By recursively applying the bound, we can trace the recurrence back to the starting point of the local epoch. Note that according to Algorithm 1, we have $\Delta_m^0=x_m^{k,h,0}-x^{k,h}=0$. Thus, we derive

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\Delta_m^i\right\|^2 \leq & \frac{1}{M}\sum_{m=1}^{M}\left[3\gamma^2l(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+3\gamma^2l\delta_2+\gamma^2\sigma^2\right]\sum_{i=0}^{l-1}\left(1+\frac{2}{l-1}\right)^i\\ \leq & \left[3\gamma^2l(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+3\gamma^2l\delta_2+\gamma^2\sigma^2\right]\sum_{i=0}^{l-1}\left(1+\frac{2}{l-1}\right)^i.\end{aligned}$$

Summing series $\sum_{i=0}^{l-1}\left(1+\frac{2}{l-1}\right)^i$ yields

$$\sum_{i=0}^{l-1}\left(1+\frac{2}{l-1}\right)^i=\frac{\left(1+\frac{2}{l-1}\right)^l-1}{\frac{2}{l-1}}=\frac{l-1}{2}\left(\left(1+\frac{2}{l-1}\right)^l-1\right)\leq\frac{l-1}{2}(e^2\cdot3-1)\leq11l.$$

Using this estimate and averaging over all local steps, we obtain the desired bound. □

Next, we need to quantify the discrepancy between the local gradients and the gradient on the server.

Lemma 6. *Suppose Assumptions 1, 3, and 4 hold. If* $\gamma\leq\frac{1}{3\cdot\sqrt{33(\delta_1+1)}\cdot lL}$*, then for Algorithm 1 we have:*

$$\mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right]\leq4(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+4\delta_2+\frac{\sigma^2}{l}.$$

Proof. We begin by expanding the squared norm:

$$\mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right] \leq \frac{3}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})-\nabla f_m(x^{k,h})\right\|^2 + \frac{3}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f_m(x^{k,h})-\nabla f(x^{k,h})\right\|^2 + \frac{3}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f(x^{k,h})\right\|^2.$$

Using the L-smoothness of f_m 1 and Assumption 4, we have

$$\mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right] \leq \frac{3L^2}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|x_m^{k,h,i}-x^{k,h}\right\|^2 + 3(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2 + 3\delta_2.$$

To complete the proof, we apply Lemma 5 and choose $\gamma \leq \frac{1}{3\cdot\sqrt{33(\delta_1+1)}\cdot lL}$. Note that this choice of γ allows us to use the result of Lemma 5 correctly, since $\gamma \leq \frac{1}{3\cdot\sqrt{33(\delta_1+1)}\cdot lL} \leq \frac{1}{\sqrt{3}\cdot lL}$. This yields the following bound:

$$\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right] \\
&\quad\leq 3L^2\left[33\gamma^2 l^2(\delta_1+1)\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f(x^{k,h})\right\|^2 + 33\gamma^2 l^2\delta_2 + 11\gamma^2 l\sigma^2\right] \\
&\qquad + 3(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2 + 3\delta_2 \\
&\quad\leq 4(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2 + 4\delta_2 + \frac{\sigma^2}{l}.
\end{aligned}$$

This finishes the proof. □

The final and most complex lemma in this section shows how the surrogate gradient behaves over the entire local epoch.

Lemma 7 (Lemma 1)**.** *Suppose Assumptions 1, 3, and 4 hold. If* $\gamma \leq \frac{1}{3\cdot\sqrt{33(\delta_1+1)}\cdot lL}$*, then for Algorithm 1 we have:*

$$\mathbb{E}_{H^k}\mathbb{E}_{\eta_m^{k,0}}\mathbb{E}_{\xi_m^{k,0}}\ldots\mathbb{E}_{\eta_m^{k,H^k-1}}\mathbb{E}_{\xi_m^{k,H^k-1}}\left\|g^{k,H^k}\right\|^2 \leq 24(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2 q_{min}}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2 + 48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2 q_{min}} + 24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2 q_{min}}.$$

Proof. We start the proof by employing Algorithm 1 to recursively step back to the previous iteration of epoch H^k. Substituting Lines 16, 17 and 27 of Algorithm 1, we estimate $g^{k,h+1}$ as follows:

$$\begin{aligned}
\left\|g^{k,h+1}\right\|^2 &= \left\|\sum_{m=1}^{M} g_m^{k,h+1}\right\|^2 = \left\|\sum_{m=1}^{M} g_m^{k,h} + \sum_{m=1}^{M}(1-\theta)\left(\frac{1}{M}-\hat{\pi}_m^k\right)\nabla_m^{k,h}\right\|^2 \\
&= \left\|g^{k,h} + \sum_{m=1}^{M}(1-\theta)\left(\frac{1}{M}-\hat{\pi}_m^k\right)\left(\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right)\right\|^2.
\end{aligned}$$

Next, applying Young's inequality yields

$$\begin{aligned}
\left\|g^{k,h+1}\right\|^2 \leq &(1+c)\left\|g^{k,h}\right\|^2 \cdot \\
&+\left(1+\frac{1}{c}\right)(1-\theta)^2\left\|\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\left(\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right)\right\|^2. \quad (5)
\end{aligned}$$

The main goal is to estimate the term 5. Due to the independence of $\xi_m^{k,h,i}$, we separate it into two components:

$$\mathbb{E}_{\xi_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2$$

$$\leq \mathbb{E}_{\xi_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\left(\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right)\right\|^2 \tag{6}$$

$$+\,\mathbb{E}_{\xi_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i})\right\|^2. \tag{7}$$

Let us begin the estimation with the first term 6. To bound it, we apply the Cauchy-Schwarz inequality.

$$\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\left(\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right)\right\|^2\\
&\leq M\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\sum_{i_1,i_2=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\Bigg[\Bigg\langle\frac{\eta_m^{k,h,i_1}}{q_m}\left(\nabla f_m(x_m^{k,h,i_1},\xi_m^{k,h,i_1})-\nabla f_m(x_m^{k,h,i_1})\right),\\
&\quad\frac{\eta_m^{k,h,i_2}}{q_m}\left(\nabla f_m(x_m^{k,h,i_2},\xi_m^{k,h,i_2})-\nabla f_m(x_m^{k,h,i_2})\right)\Bigg\rangle\Bigg]\\
&\leq M\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right\|^2\\
&\leq \sigma^2 M\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2,
\end{aligned}$$

since $\xi_m^{k,h,i}$ are independent. Next, we consider $\eta_m^{k,h,i}$. Note that $\eta_m^{k,h,i}\sim Be(q_m)$, so $\mathbb{E}_{\eta_m^{k,h,i}}\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\leq\frac{1}{q_m}$. Thus, taking the expectation with respect to $\eta_m^{k,h}$ yields

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2\leq\sigma^2\frac{l}{q_{min}}M\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2.$$

The regularized simplex constraint given by the intersection of the standard simplex and the inequalities $\mathbf{e}_m^\top\pi-\frac{\alpha}{M}\leq 0$ ensures that no weight falls below $\frac{\alpha}{M}$. The worst-case deviation from the uniform distribution $1/M$ occurs when as many weights as possible are set to the lower bound, i.e., $\lfloor\frac{M}{\alpha}\rfloor$ coefficients equal $\frac{\alpha}{M}$ and the rest are zero. Noting that in such case we have $\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\leq\frac{\alpha}{M}$, we obtain:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2\leq\frac{l\alpha}{q_{min}}\sigma^2.$$

We have finished estimating the first of the two terms needed to bound 5. To complete the estimation, we also need to bound 7. The idea is exactly the same as for 6. We begin with the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i})\right\|^2\\
&\leq l\sum_{i=0}^{l-1}\left(\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right).
\end{aligned}$$

Next, we take the additional expectation. Recall the inequality $\mathbb{E}_{\eta_m^{k,h,i}}\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2 \leq \frac{1}{q_m}$ from the discussion above. Applying it, we obtain:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i})\right\|^2 \leq \frac{l}{q_{min}}\sum_{i=0}^{l-1}\left(\sum_{m=1}^{M}\left(\frac{1}{M}-\hat{\pi}_m^k\right)^2\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right).$$

The next step is to use the constraints on the weights to get the desired bound.

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i})\right\|^2 \leq \frac{l^2\alpha}{q_{min}}\frac{1}{lM}\sum_{i=0}^{l-1}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2.$$

To complete the estimation of this term, it remains to apply Lemma 6. From this, we can immediately write the final bound for the term 5:

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left(\frac{1}{M}-\hat{\pi}_m^k\right)\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2 &\leq \frac{l\alpha}{q_{min}}\sigma^2+\frac{l^2\alpha}{q_{min}}\left(4(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+4\delta_2+\frac{\sigma^2}{l}\right)\\ &\leq \frac{4l^2\alpha}{q_{min}}(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+\frac{4l^2\alpha}{q_{min}}\delta_2+\frac{2l\alpha}{q_{min}}\sigma^2.\end{aligned}$$

With the term 5 now fully bounded, we return to estimating $\|g^{k,h+1}\|^2$. Note that $g^{k,0}=0$ by Line 6 of Algorithm 1. Moreover, we set $c=\frac{p}{2}$, which yields $1+1/c=1+2/p\leq 3/p$ for $p\in(0,1]$. Taking additional expectations, we obtain:

$$\begin{aligned}\mathbb{E}_{\eta_m^{k,0}}\mathbb{E}_{\xi_m^{k,0}}\ldots\mathbb{E}_{\eta_m^{k,h}}\mathbb{E}_{\xi_m^{k,h}}\left\|g^{k,h+1}\right\|^2 \leq{}& 12(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{pq_{min}}\sum_{i=0}^{h}\left\|\nabla f(x^{k,i})\right\|^2\left(1+\frac{p}{2}\right)^{h-i}\\ &+12(1-\theta)^2\frac{l^2\alpha\delta_2}{pq_{min}}\sum_{i=0}^{h}\left(1+\frac{p}{2}\right)^{h-i}\\ &+6(1-\theta)^2\frac{l\alpha\sigma^2}{pq_{min}}\sum_{i=0}^{h}\left(1+\frac{p}{2}\right)^{h-i}.\end{aligned}$$

Finally, since $H^k\sim\mathrm{Geom}(p)$, we may apply Lemmas 2 and 3. Consequently, we obtain:

$$\begin{aligned}&\mathbb{E}_{H^k}\mathbb{E}_{\eta_m^{k,0}}\mathbb{E}_{\xi_m^{k,0}}\ldots\mathbb{E}_{\eta_m^{k,H^k-1}}\mathbb{E}_{\xi_m^{k,H^k-1}}\left\|g^{k,H^k}\right\|^2\\ &\quad=\sum_{j\geq 0}p(1-p)^j\mathbb{E}_{\eta_m^{k,0}}\mathbb{E}_{\xi_m^{k,0}}\ldots\mathbb{E}_{\eta_m^{k,j-1}}\mathbb{E}_{\xi_m^{k,j-1}}\left\|g^{k,j}\right\|^2\\ &\quad\leq 24(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2\\ &\qquad+48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}\\ &\qquad+24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}}.\end{aligned}$$

□

C PPBC WITH LOCAL STEPS, NON-CONVEX

Theorem 2 (Theorem 1)**.** *Suppose Assumptions 1, 2, 3, and 4 hold. Then for Algorithm 1 with parameters satisfying*

$$\gamma \leq \frac{p \cdot q_{min}}{33Ll\alpha(\delta_1+1)}, \qquad \theta \leq \frac{\gamma L p^2}{2}$$

we have:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 \leq \frac{16}{\gamma K l}\mathbb{E}\left[f(x^{0,0})-f(x^*)\right] \\ + 384L\gamma\frac{\alpha\delta_2}{pq_{min}} + 24L\gamma\frac{l\delta_2}{pq_{min}} + 192L^2\gamma^2\frac{\alpha\delta_2 l^2}{p^3 q_{min}} \\ + 192L\gamma\frac{\alpha\sigma^2}{lpq_{min}} + 16L\gamma\frac{\sigma^2}{pq_{min}} + 96L^2\gamma^2\frac{l\alpha\sigma^2}{p^3 q_{min}}.\end{aligned}$$

Proof. The proof requires a simple trick. Instead of writing the descent in terms of $x^{k,h}$, we write it using the virtual sequence $\tilde{x}^{k,h}$. It is defined as follows:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma\sum_{m=1}^{M} g_m^{k,h} = x^{k,h} - \gamma g^{k,h}, \tag{8}$$

where $g^{k,h} = \sum\limits_{m=1}^{M} g_m^{k,h}$ is the aggregated gradient.

This allows us to deal with the gradients from all clients. Note that from the Algorithm 1 (Lines 6 and 7), we have $\tilde{x}^{k,H^k} = x^{k,H^k} - \gamma g^{k,H^k} = x^{k+1,0}$ and $\tilde{x}^{k,0} = x^{k,0}$, which enables us to return to the original $x^{k,h}$ variables at the end. We begin the proof by deriving the update rule for the virtual sequence.

$$\begin{aligned}\tilde{x}^{k,h} &= x^{k,h} - \gamma\sum_{m=1}^{M} g_m^{k,h} = x^{k,h-1} - \gamma\left((1-\theta)\sum_{m=1}^{M}\hat{\pi}_m^k\nabla_m^{k,h-1} + \theta g^{k-1,H^{k-1}}\right) \\ &\quad - \gamma\sum_{m=1}^{M}\left(g_m^{k,h-1} + (1-\theta)\left(\frac{1}{M}-\hat{\pi}_m^k\right)\nabla_m^{k,h-1}\right) \\ &= x^{k,h-1} - \gamma\sum_{m=1}^{M} g_m^{k,h-1} - \gamma\left((1-\theta)\sum_{m=1}^{M}\frac{1}{M}\nabla_m^{k,h-1} + \theta g^{k-1,H^{k-1}}\right) \\ &= \tilde{x}^{k,h-1} - \gamma\left((1-\theta)\sum_{m=1}^{M}\frac{1}{M}\nabla_m^{k,h-1} + \theta g^{k-1,H^{k-1}}\right).\end{aligned}$$

Next, we need to write the descent along the virtual sequence. To this end, we use the L-smoothness property and the update rule.

$$\begin{aligned}f(\tilde{x}^{k,h+1}) &\leq f(\tilde{x}^{k,h}) + \left\langle\nabla f(\tilde{x}^{k,h}), \tilde{x}^{k,h+1}-\tilde{x}^{k,h}\right\rangle + \frac{L}{2}\left\|\tilde{x}^{k,h+1}-\tilde{x}^{k,h}\right\|^2 \\ &\leq f(\tilde{x}^{k,h}) - \gamma\theta\left\langle\nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}}\right\rangle \qquad (9) \\ &\quad - \gamma(1-\theta)\left\langle\nabla f(\tilde{x}^{k,h}), \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle \qquad (10) \\ &\quad + \frac{L\gamma^2(1-\theta)}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2 \qquad (11) \\ &\quad + \frac{L\gamma^2\theta}{2}\left\|g^{k-1,H^{k-1}}\right\|^2.\end{aligned}$$

We estimate the terms in 9,10 and 11 separately. Let us start with the term 9. Here, we simply apply Young's inequality.

$$-\gamma\theta\left\langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}}\right\rangle \leq \frac{\gamma\theta l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2 + \frac{\gamma\theta}{2l}\left\|g^{k-1,H^{k-1}}\right\|^2.$$

Next, we turn to estimating the term 10. To this end, we take the expectation over $\eta_m^{k,h}$. Since $\eta_m^{k,h,i} \sim \mathrm{Be}(q_m)$ are independent, this step removes the stochasticity of client availability and yields

$$\begin{aligned}\mathbb{E}_{\eta_m^{k,h}}\Bigg[-\gamma(1-\theta)\Bigg\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\Bigg\rangle\Bigg] \\ = -\gamma(1-\theta)\underbrace{\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle}_{(*)}.\end{aligned}$$

The key challenge is to bound the following inner product term (*). We now take expectation with respect to $\xi_m^{k,h}$ and simultaneously add and subtract $l\cdot\nabla f(x^{k,h})$:

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k,h}}\Big[-\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Big] \\ &\quad = -l\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h})\right\rangle - \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i}) - \nabla f(x^{k,h})\right\rangle.\end{aligned}$$

After this, we can transform the inner products to obtain an alternative form of the expression.

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k,h}}\Big[-\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Big] && (12)\\ &\quad = -\frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2 - \frac{l}{2}\left\|\nabla f(x^{k,h})\right\|^2 + \frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\right\|^2 \\ &\qquad - \mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i}) - \nabla f_m(x_m^{k,h,i})\right\rangle\right] && (13)\\ &\qquad - \mathbb{E}_{\xi_m^{k,h}}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i}) - \nabla f_m(x^{k,h})\right\rangle\right] \\ &\qquad - \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\right\rangle. && (14)\end{aligned}$$

Note that the term 13 equals zero. Also, in view of Assumption 4, the term 14 equals zero. Then we can rewrite12 as follows:

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k,h}}\Bigg[-\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Bigg] \\ &\quad = -\frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2 - \frac{l}{2}\left\|\nabla f(x^{k,h})\right\|^2 + \frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\right\|^2 \\ &\qquad - \frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\langle \nabla f(\tilde{x}^{k,h}), \nabla f_m(x_m^{k,h,i}) - \nabla f_m(x^{k,h})\right\rangle.\end{aligned}$$

Using both Young's inequality and L-smoothness, we get

$$
\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\Bigg[-\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle\nabla f(\tilde{x}^{k,h}),\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Bigg]\\
&\quad\leq -\frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{l}{2}\left\|\nabla f(x^{k,h})\right\|^2+\frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h})-\nabla f(x^{k,h})\right\|^2+\frac{l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2\\
&\qquad+\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})-\nabla f_m(x^{k,h})\right\|^2\\
&\quad\leq -\frac{l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{l}{2}\left\|\nabla f(x^{k,h})\right\|^2+\frac{L^2 l}{2}\left\|\tilde{x}^{k,h}-x^{k,h}\right\|^2+\frac{l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2\\
&\qquad+L^2\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|x_m^{k,h,i}-x^{k,h}\right\|^2.
\end{aligned}
$$

Recall that by definition of the virtual sequence: $\left\|\tilde{x}^{k,h}-x^{k,h}\right\|^2=\left\|\gamma g^{k,h}\right\|^2$. Hence, this expression can be rewritten as:

$$
\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\Bigg[-\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\left\langle\nabla f(\tilde{x}^{k,h}),\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Bigg]\\
&\qquad\leq -\frac{l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{l}{2}\left\|\nabla f(x^{k,h})\right\|^2+\frac{L^2\gamma^2 l}{2}\left\|g^{k,h}\right\|^2\\
&\qquad\quad+l\cdot\frac{L^2}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|x_m^{k,h,i}-x^{k,h}\right\|^2.
\end{aligned}
$$

To bound the last term, we apply Lemma 5. Thus, the final bound for term 10 is as follows:

$$
\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\Bigg[-\gamma(1-\theta)\left\langle\nabla f(\tilde{x}^{k,h}),\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\rangle\Bigg]\\
&\qquad\leq-\frac{\gamma(1-\theta)l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{\gamma(1-\theta)l}{2}\left\|\nabla f(x^{k,h})\right\|^2+\frac{L^2\gamma^3(1-\theta)l}{2}\left\|g^{k,h}\right\|^2\\
&\qquad\quad+33L^2\gamma^3l^3(1-\theta)(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+33L^2\gamma^3l^3(1-\theta)\delta_2\\
&\qquad\quad+11L^2\gamma^3l^2(1-\theta)\sigma^2\\
&\qquad=-\frac{\gamma(1-\theta)l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{\gamma(1-\theta)l}{2}\left(1-66L^2\gamma^2l^2(\delta_1+1)\right)\left\|\nabla f(x^{k,h})\right\|^2\\
&\qquad\quad+\frac{L^2\gamma^3(1-\theta)l}{2}\left\|g^{k,h}\right\|^2+33L^2\gamma^3l^3(1-\theta)\delta_2+11L^2\gamma^3l^2(1-\theta)\sigma^2.
\end{aligned}
$$

It remains to estimate the term 11. We take the expectation over $\eta_m^{k,h}$ and $\xi_m^{k,h}$ and apply the Cauchy-Schwarz inequality. Due to the independence of the random variables $\xi_m^{k,h,i}$, we obtain:

$$
\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left[\frac{L\gamma^2(1-\theta)}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2\right]\\
&\quad\leq\frac{L\gamma^2(1-\theta)}{2}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\left(\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right)\right\|^2 \qquad (15)\\
&\qquad+\frac{L\gamma^2(1-\theta)}{2}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i})\right\|^2. \qquad (16)
\end{aligned}
$$

To estimate the term 15, we apply the following trick. If we expand the norm via the inner product, then, due to independence of $\xi_m^{k,h,i}$, we can obtain the following equality:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\left(\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right)\right\|^2$$
$$=\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\sum_{i=0}^{l-1}\left\|\frac{\eta_m^{k,h,i}}{q_m}\left(\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right)\right\|^2.$$

As for the term 16, we simply apply the Cauchy-Schwarz inequality to the sum over l. Together, these two considerations give:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left[\frac{L\gamma^2(1-\theta)}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2\right]$$
$$\leq\frac{L\gamma^2(1-\theta)}{2}\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left[\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\left\|\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})-\nabla f_m(x_m^{k,h,i})\right\|^2\right]$$
$$+\frac{L\gamma^2(1-\theta)}{2}l^2\cdot\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left[\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\left\|\nabla f_m(x_m^{k,h,i})\right\|^2\right].$$

Applying the inequality $\mathbb{E}_{\eta_m^{k,h,i}}\left(\frac{\eta_m^{k,h,i}}{q_m}\right)^2\leq\frac{1}{q_m}$ yields

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left[\frac{L\gamma^2(1-\theta)}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2\right]$$
$$\leq\frac{L\gamma^2(1-\theta)l}{2q_{min}}\sigma^2+\frac{L\gamma^2(1-\theta)}{2q_{min}}l^2\frac{1}{lM}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\mathbb{E}_{\xi_m^{k,h}}\left\|\nabla f_m(x_m^{k,h,i})\right\|^2.$$

All that remains is to apply Lemma 6, then we obtain:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\frac{L\gamma^2(1-\theta)}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\sum_{i=0}^{l-1}\frac{\eta_m^{k,h,i}}{q_m}\nabla f_m(x_m^{k,h,i},\xi_m^{k,h,i})\right\|^2$$
$$\leq\frac{2L\gamma^2(1-\theta)l^2}{q_{min}}(\delta_1+1)\left\|f(x^{k,h})\right\|^2+\frac{2L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2+\frac{L\gamma^2(1-\theta)l}{q_{min}}\sigma^2.$$

We are now prepared to proceed with the recursive descent expansion on the virtual sequence and derive a bound for $f(\tilde{x}^{k,h+1})$. Let us write down the bound, taking into account the inequalities obtained above.

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}f(\tilde{x}^{k,h+1})\leq\ & f(\tilde{x}^{k,h})+\frac{\gamma\theta l}{2}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2+\frac{\gamma\theta}{2l}\left\|g^{k-1,H^{k-1}}\right\|^2\\&-\frac{\gamma(1-\theta)l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2-\frac{\gamma(1-\theta)l}{2}\left(1-66L^2\gamma^2l^2(\delta_1+1)\right)\left\|\nabla f(x^{k,h})\right\|^2\\&+\frac{L^2\gamma^3(1-\theta)l}{2}\left\|g^{k,h}\right\|^2+33L^2\gamma^3l^3(1-\theta)\delta_2+11L^2\gamma^3l^2(1-\theta)\sigma^2\\&+\frac{2L\gamma^2(1-\theta)l^2}{q_{min}}(\delta_1+1)\left\|\nabla f(x^{k,h})\right\|^2+\frac{2L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2+\frac{L\gamma^2(1-\theta)l}{q_{min}}\sigma^2\\&+\frac{L\gamma^2\theta}{2}\left\|g^{k-1,H^{k-1}}\right\|^2\\=\ & f(\tilde{x}^{k,h})-\frac{\gamma(1-3\theta)l}{4}\left\|\nabla f(\tilde{x}^{k,h})\right\|^2+\frac{\gamma\theta(1+L\gamma l)}{2l}\left\|g^{k-1,H^{k-1}}\right\|^2\\&-\frac{\gamma(1-\theta)l}{2}\left(1-66L^2\gamma^2l^2(\delta_1+1)-\frac{4Ll\gamma(\delta_1+1)}{q_{min}}\right)\left\|\nabla f(x^{k,h})\right\|^2\\&+\frac{L^2\gamma^3(1-\theta)l}{2}\left\|g^{k,h}\right\|^2+\frac{L\gamma^2(1-\theta)l^2}{q_{min}}(2+33L\gamma lq_{min})\delta_2\\&+\frac{L\gamma^2(1-\theta)l}{q_{min}}(1+11L\gamma lq_{min})\sigma^2.\end{aligned}$$

To continue, we now choose the parameters γ and θ. We select $\gamma \leq \frac{q_{\min}}{33Ll(\delta_1+1)}$. Then we have:

$$1 - 66L^2\gamma^2 l^2(\delta_1+1) - \frac{4L\gamma l(\delta_1+1)}{q_{\min}} \geq \frac{1}{2},$$
$$1 + L\gamma l \leq 2, \qquad 2 + 33L\gamma l q_{\min} \leq 3, \qquad 1 + 11L\gamma l q_{\min} \leq 2.$$

Additionally, choosing $\theta \leq \frac{1}{3}$, we obtain:

$$\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}} f(\tilde{x}^{k,h+1}) \leq f(\tilde{x}^{k,h}) + \frac{\gamma\theta}{l}\left\|g^{k-1,H^{k-1}}\right\|^2 - \frac{\gamma(1-\theta)l}{4}\left\|\nabla f(x^{k,h})\right\|^2 + \frac{L^2\gamma^3(1-\theta)l}{2}\left\|g^{k,h}\right\|^2 + \frac{3L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2 + \frac{2L\gamma^2(1-\theta)l}{q_{min}}\sigma^2.$$

Now we set $h = H^k - 1$ and take additional expectations:

$$\begin{aligned}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k})\right] &\leq \mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k-1})\right] \\ &+ \frac{\gamma\theta}{l}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\ &- \frac{\gamma(1-\theta)l}{4}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|\nabla f(x^{k,H^k-1})\right\|^2 \\ &+ \frac{L^2\gamma^3(1-\theta)l}{2}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|g^{k,H^k-1}\right\|^2 \\ &+ \frac{3L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2 + \frac{2L\gamma^2(1-\theta)l}{q_{min}}\sigma^2.\end{aligned}$$

Taking expectations with respect to $H^k - 1$ and H^k and applying Lemma 4, we obtain:

$$\begin{aligned}&\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k})\right] \\ &\leq (1-p)\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k})\right] \\ &+ p\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left[f(\tilde{x}^{k,0})\right] \\ &+ \frac{\gamma\theta}{l}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \qquad (17) \\ &- (1-p)\frac{\gamma(1-\theta)l}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|\nabla f(x^{k,H^k})\right\|^2 \\ &- p\frac{\gamma(1-\theta)l}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\ldots\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\ &+ (1-p)\frac{L^2\gamma^3(1-\theta)l}{2} \\ &\cdot \mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k}}\left\|g^{k,H^k}\right\|^2 \qquad (18) \\ &+ p\frac{L^2\gamma^3(1-\theta)l}{2}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|g^{k,0}\right\|^2 \qquad (19) \\ &+ \frac{3L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2 + \frac{2L\gamma^2(1-\theta)l}{q_{min}}\sigma^2.\end{aligned}$$

We now examine the terms 17, 18, and 19 one by one. Due to the update rules of the algorithm, the term 19 vanishes. It remains to estimate 17 and 18. We begin with 17 and apply Lemma 1. Consequently, we obtain:

$$\begin{aligned}&\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\ &\leq 24(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2 q_{min}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k-1,H^{k-1}})\right\|^2 \\ &+ 48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2 q_{min}} + 24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2 q_{min}}.\end{aligned}$$

Next, we apply the triangle inequality.

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\
&\quad\leq 48(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k-1,H^{k-1}})-\nabla f(\tilde{x}^{k-1,H^{k-1}})\right\|^2 \\
&\qquad+48(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(\tilde{x}^{k-1,H^{k-1}})\right\|^2 \\
&\qquad+48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}+24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}}.
\end{aligned}$$

To proceed, we need three facts: L-smoothness, the definition of the virtual sequence 8, and the equality $\tilde{x}^{k-1,H^{k-1}}=x^{k,0}$, which follows directly from the algorithm's update rules.

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\
&\qquad\leq 48L^2(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|x^{k-1,H^{k-1}}-\tilde{x}^{k-1,H^{k-1}}\right\|^2 \\
&\qquad\quad+48(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
&\qquad\quad+48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}+24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}} \\
&\qquad=48L^2\gamma^2(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\
&\qquad\quad+48(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
&\qquad\quad+48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}+24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}}.
\end{aligned}$$

Let us once again impose constraints on γ. We choose γ satisfying $\gamma\leq\frac{\sqrt{q_{\min}}}{2Ll\sqrt{48\alpha(\delta_1+1)}}$. Moreover, we take additional expectations $\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^k-1}}\mathbb{E}_{\eta_m^{k-1,H^k-1}}[\cdot]$. Consequently, one can obtain:

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^k-1}}\mathbb{E}_{\eta_m^{k-1,H^k-1}}\left\|g^{k-1,H^{k-1}}\right\|^2 \\
&\quad\leq 96(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^k-1}}\mathbb{E}_{\eta_m^{k-1,H^k-1}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
&\qquad+96(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}+48(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}}.
\end{aligned}$$

As for the term 18, it suffices to apply Lemma 1, which immediately yields

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k}}\left\|g^{k,H^k}\right\|^2 \\
&\qquad\leq 24(1-\theta)^2\frac{l^2\alpha(\delta_1+1)}{p^2q_{min}} \\
&\qquad\quad\cdot\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\ldots\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
&\qquad\quad+48(1-\theta)^2\frac{l^2\alpha\delta_2}{p^2q_{min}}+24(1-\theta)^2\frac{l\alpha\sigma^2}{p^2q_{min}}.
\end{aligned}$$

Putting everything together and observing that, by definition, $\tilde{x}^{k,H^k}=x^{k,H^k}-\gamma g^{k,H^k}=x^{k+1,0}$ and $\tilde{x}^{k,0}=x^{k,0}$, we obtain:

$$
\begin{aligned}
&p\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k})\right] \\
&\quad\le p\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left[f(\tilde{x}^{k,0})\right] \\
&\qquad+96\gamma\theta(1-\theta)^2\frac{l\alpha(\delta_1+1)}{p^2q_{min}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1}}\mathbb{E}_{\eta_m^{k-1,H^k-1}}\mathbb{E}_{H^{k-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
&\qquad+96\gamma\theta(1-\theta)^2\frac{l\alpha\delta_2}{p^2q_{min}}+48\gamma\theta(1-\theta)^2\frac{\alpha\sigma^2}{p^2q_{min}} \\
&\qquad-(1-p)\frac{\gamma(1-\theta)l}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
&\qquad-p\frac{\gamma(1-\theta)l}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
&\qquad+12(1-p)L^2\gamma^3(1-\theta)^3\frac{l^3\alpha(\delta_1+1)}{p^2q_{min}} \\
&\qquad\cdot\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
&\qquad+24(1-p)L^2\gamma^3(1-\theta)^3\frac{l^3\alpha\delta_2}{p^2q_{min}}+12(1-p)L^2\gamma^3(1-\theta)^3\frac{l^2\alpha\sigma^2}{p^2q_{min}} \\
&\qquad+\frac{3L\gamma^2(1-\theta)l^2}{q_{min}}\delta_2+\frac{2L\gamma^2(1-\theta)l}{q_{min}}\sigma^2.
\end{aligned}
$$

Taking the full expectation, applying the law of total expectation, and rearranging terms yields:

$$
\begin{aligned}
\frac{4}{\gamma(1-\theta)l}\mathbb{E}\left[f(\tilde{x}^{k,H^k})\right]\le{}&\frac{4}{\gamma(1-\theta)l}\mathbb{E}\left[f(\tilde{x}^{k,0})\right] \\
&-\left(1-\frac{384\theta(1-\theta)}{p^3q_{min}}\alpha(\delta_1+1)\right)\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 &(20)\\
&-\frac{1-p}{p}\left(1-\frac{48L^2\gamma^2(1-\theta)^2l^2}{p^2q_{min}}\alpha(\delta_1+1)\right)\mathbb{E}\left\|\nabla f(x^{k,H^k})\right\|^2 &(21)\\
&+384\theta(1-\theta)\frac{\alpha\delta_2}{p^3q_{min}}+192\theta(1-\theta)\frac{\alpha\sigma^2}{lp^3q_{min}} \\
&+12L\gamma\frac{l\delta_2}{pq_{min}}+8L\gamma\frac{\sigma^2}{pq_{min}} \\
&+96(1-p)L^2\gamma^2(1-\theta)^2l^2\frac{\alpha\delta_2}{p^3q_{min}}+48(1-p)L^2\gamma^2(1-\theta)^2\frac{l\alpha\sigma^2}{p^3q_{min}}.
\end{aligned}
$$

To ensure the coefficients of the gradient norms are negative, we select the parameters as follows:

$$
\gamma\le\frac{p\sqrt{q_{min}}}{Ll\sqrt{192\alpha(\delta_1+1)}},\qquad\theta\le\min\left\{\frac{p\sqrt{pq_{min}}}{\sqrt{16\cdot96\alpha(\delta_1+1)}},\frac{\gamma Lp^2}{2}\right\}.
$$

Note that under this choices $\frac{1}{2}\le(1-\theta)\le1$. Thus, we obtain the following lower bounds for the coefficients:

$$
\left(1-\frac{48L^2\gamma^2(1-\theta)^2l^2}{p^2q_{min}}\alpha(\delta_1+1)\right)\ge\frac{1}{2},\qquad\left(1-\frac{384\theta(1-\theta)}{p^3q_{min}}\alpha(\delta_1+1)\right)\ge\frac{1}{2}. \tag{22}
$$

Plugging 22 into 20 and 21 and moving the gradient term to the left gives:

$$
\begin{aligned}
\frac{1}{2}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2\le{}&\frac{8}{\gamma l}\mathbb{E}\left[f(\tilde{x}^{k,0})-f(\tilde{x}^{k,H^k})\right] \\
&+192L\gamma\frac{\alpha\delta_2}{pq_{min}}+12L\gamma\frac{l\delta_2}{pq_{min}}+96L^2\gamma^2\frac{\alpha\delta_2l^2}{p^3q_{min}} \\
&+96L\gamma\frac{\alpha\sigma^2}{lpq_{min}}+8L\gamma\frac{\sigma^2}{pq_{min}}+48L^2\gamma^2\frac{l\alpha\sigma^2}{p^3q_{min}}.
\end{aligned}
$$

Summing for $k = 0, \ldots, K-1$ and dividing by K gives the final convergence bound:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 \leq \frac{16}{\gamma K l}\mathbb{E}\left[f(x^{0,0}) - f(x^*)\right] + 384L\gamma\frac{\alpha\delta_2}{pq_{min}} + 24L\gamma\frac{l\delta_2}{pq_{min}} + 192L^2\gamma^2\frac{\alpha\delta_2 l^2}{p^3 q_{min}} + 192L\gamma\frac{\alpha\sigma^2}{lpq_{min}} + 16L\gamma\frac{\sigma^2}{pq_{min}} + 96L^2\gamma^2\frac{l\alpha\sigma^2}{p^3 q_{min}}.$$

□

For completeness, we need to determine how many communication rounds are required to achieve accuracy ε.

Corollary 2 (Corollary 1)**.** *Under the conditions of Theorem 1, Algorithm 1 needs*

$$\mathcal{O}\left(\frac{\Delta L\alpha\delta_1}{pq_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lpq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{pq_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2 pq_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lpq_{min}\varepsilon^4}\right) \text{ iterations and}$$

$$\mathcal{O}\left(M\left(\frac{M}{\min\limits_k C^k}\right)^2 \cdot \left(\frac{\Delta L\alpha\delta_1}{q_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{q_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2 q_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lq_{min}\varepsilon^4}\right)\right)$$

number of devices communications to achieve ε-accuracy. Here $\Delta = f(x^{0,0}) - f(x^*)$, $\varepsilon^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2$, C^k is the number of devices participating in the k-th iteration.

Proof. We choose

$$\gamma \leq \min\left\{\frac{pq_{min}}{33Ll\alpha(\delta_1+1)}, \frac{\sqrt{pq_{min}\Delta}}{\sqrt{384L\alpha\delta_2 lK}}, \frac{\sqrt{pq_{min}\Delta}}{l\sqrt{24L\delta_2 K}}, \frac{p\sqrt[3]{\Delta q_{min}}}{l\sqrt[3]{192L^2\alpha\delta_2 K}}, \frac{\sqrt{pq_{min}\Delta}}{\sqrt{192L\alpha\sigma^2 K}}, \frac{\sqrt{pq_{min}\Delta}}{\sqrt{16L\sigma^2 lK}}, \frac{p\sqrt[3]{\Delta q_{min}}}{\sqrt[3]{96L^2\alpha\sigma^2 l^2 K}}\right\}.$$

Then we use the result of Theorem 1 and get the upper bound for the number of iterations:

$$\mathcal{O}\left(\frac{\Delta L\alpha\delta_1}{pq_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lpq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{pq_{min}\varepsilon^4} + \frac{\Delta L\alpha^{\frac{1}{2}}\delta_2^{\frac{1}{2}}}{p^{\frac{3}{2}}q_{min}^{\frac{1}{2}}\varepsilon^3} + \frac{\Delta L\alpha\sigma^2}{l^2 pq_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lpq_{min}\varepsilon^4} + \frac{\Delta L\alpha^{\frac{1}{2}}\sigma}{p^{\frac{3}{2}}q_{min}^{\frac{1}{2}}l^{\frac{1}{2}}\varepsilon^3}\right).$$

Ignoring the terms that do not contribute to the bound yields

$$\mathcal{O}\left(\frac{\Delta L\alpha\delta_1}{pq_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lpq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{pq_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2 pq_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lpq_{min}\varepsilon^4}\right).$$

Since the length of each epoch is sampled from a geometric distribution, the expected number of iterations per epoch is $1/p$. The expected number of communications with server per epoch is $C^k \cdot \frac{1}{p} + M$, where C^k is a number of devices which are ready for communication in iteration k. We have to minimize this number, so we choose $p = \frac{\min\limits_k C^k}{M}$. Consequently, we obtain the number of communications in each epoch $\mathcal{O}\left(M\frac{M}{\min\limits_k C^k}\right)$ and the total number of device communications

$$\mathcal{O}\left(M\left(\frac{M}{\min\limits_k C^k}\right)^2 \cdot \left(\frac{\Delta L\alpha\delta_1}{q_{min}\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{lq_{min}\varepsilon^4} + \frac{\Delta L\delta_2}{q_{min}\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{l^2 q_{min}\varepsilon^4} + \frac{\Delta L\sigma^2}{lq_{min}\varepsilon^4}\right)\right).$$

□

D EXPERIMENTS

This appendix supplements the experimental setup and results presented in the main text with additional details that were omitted for brevity. Unless stated otherwise, all experiments reported here follow the same protocols, datasets, and hyperparameter configurations described in Section 6. Here, besides other additional information, we present extended tables and supplementary plots that complement the aggregate results reported earlier. These materials are intended to give readers a more complete picture of our findings and to facilitate the replication of our work.

D.1 EXPERIMENT RESULTS UNDER PARTIAL PARTICIPATION

In this subsection, we present a complementary set of results. While the tables in the main part reported the best accuracy achieved after 100 epochs, the tables below report the number of epochs each algorithm required to reach a given accuracy threshold, illustrating the speed at which each method converges. The accompanying plots visualize these speed comparisons.

Algorithm	PoC	FedCBS
PP-EFLS	68	32
FedAvg	100+	53
FedProx	100+	45

Table 9: Number of epochs to achieve accuracy 0.25; $\alpha = 0.1$, LS1

Algorithm	PoC	FedCBS
PP-EFLS	77	36
FedAvg	100+	73
FedProx	86	65

Table 10: Number of epochs to achieve accuracy 0.4; $\alpha = 0.1$, LS5

Algorithm	PoC	FedCBS
PP-EFLS	38	40
FedAvg	100+	71
FedProx	100+	71

Table 11: Number of epochs to achieve accuracy 0.5; $\alpha = 0.1$, LS10

Algorithm	PoC	FedCBS
PP-EFLS	14	11
FedAvg	49	24
FedProx	40	26

Table 12: Number of epochs to achieve accuracy 0.3; $\alpha = 5$, LS1

Algorithm	PoC	FedCBS
PP-EFLS	32	29
FedAvg	73	60
FedProx	73	60

Table 13: Number of epochs to achieve accuracy 0.5; $\alpha = 5$, LS5

Algorithm	PoC	FedCBS
PP-EFLS	30	28
FedAvg	69	60
FedProx	73	60

Table 14: Number of epochs to achieve accuracy 0.6; $\alpha = 5$, LS10

Algorithm	PoC	FedCBS
PP-EFLS	10	13
FedAvg	36	27
FedProx	32	32

Table 15: Number of epochs to achieve accuracy 0.3; $\alpha = 10000$, LS1

Algorithm	PoC	FedCBS
PP-EFLS	29	30
FedAvg	63	69
FedProx	71	66

Table 16: Number of epochs to achieve accuracy 0.5; $\alpha = 10000$, LS5

Algorithm	PoC	FedCBS
PP-EFLS	28	26
FedAvg	73	60
FedProx	71	65

Table 17: Number of epochs to achieve accuracy 0.6; $\alpha = 10000$, LS10

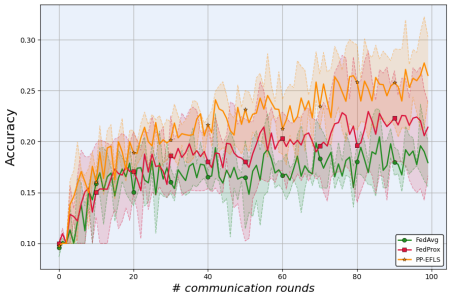


Figure 1: PoC; $\alpha = 0.1$; LS=1

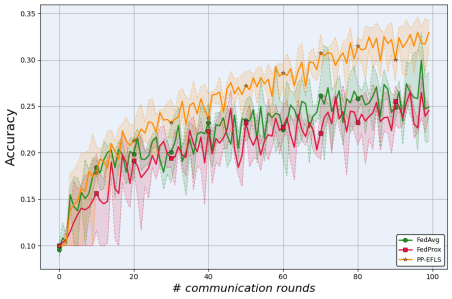


Figure 2: FedCBS; $\alpha = 0.1$; LS=1

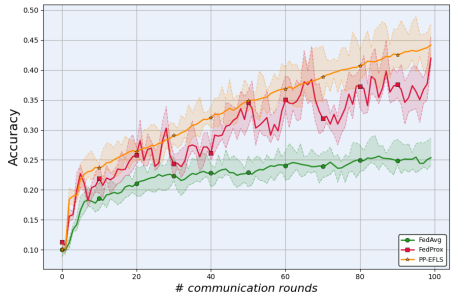


Figure 3: PoC; $\alpha = 0.1$; LS=5

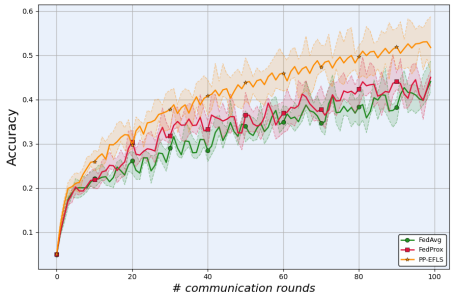


Figure 4: FedCBS; $\alpha = 0.1$; LS=5

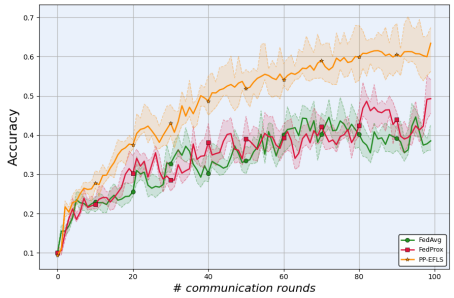


Figure 5: PoC; $\alpha = 0.1$; LS=10

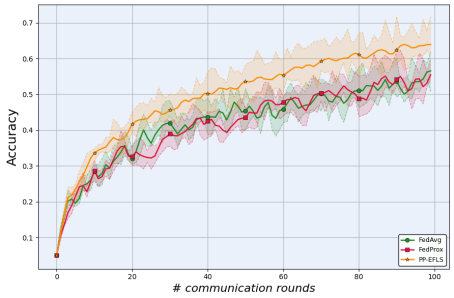


Figure 6: FedCBS; $\alpha = 0.1$; LS=10

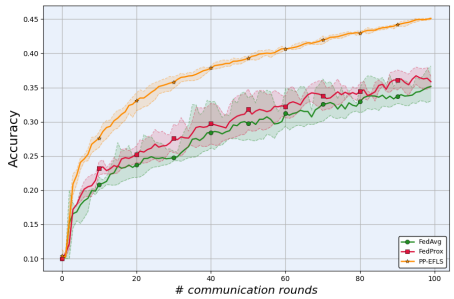


Figure 7: PoC; $\alpha = 5$; LS=1

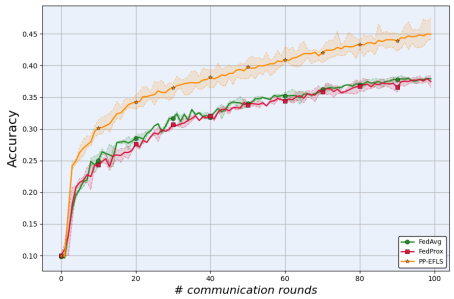


Figure 8: FedCBS; $\alpha = 5$; LS=1

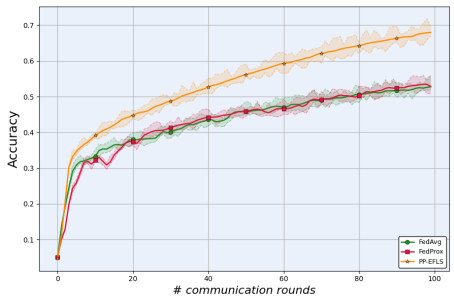


Figure 9: PoC; $\alpha = 5$; LS=5

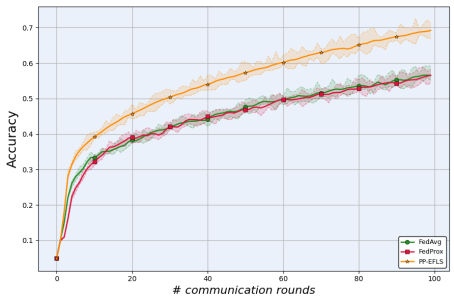


Figure 10: FedCBS; $\alpha = 5$; LS=5

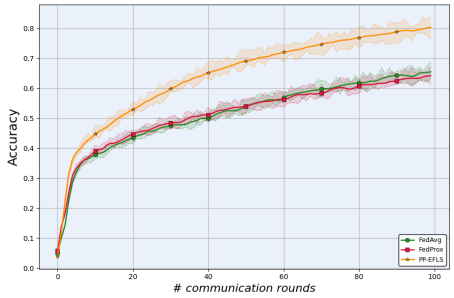


Figure 11: PoC; $\alpha = 5$; LS=10

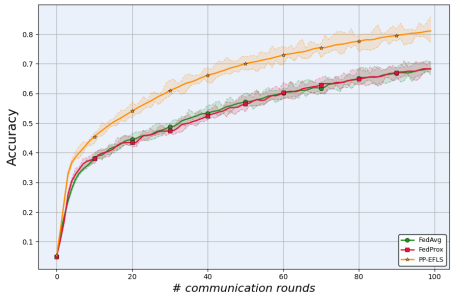


Figure 12: FedCBS; $\alpha = 5$; LS=10

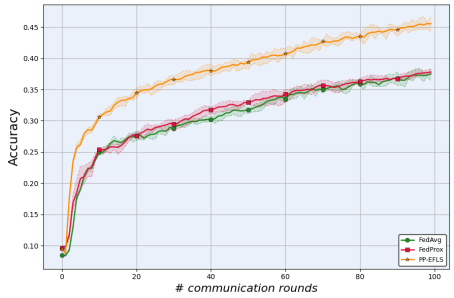


Figure 13: PoC; $\alpha = 10000$; LS=1

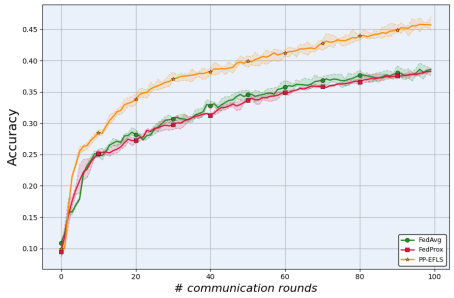


Figure 14: FedCBS; $\alpha = 10000$; LS=1

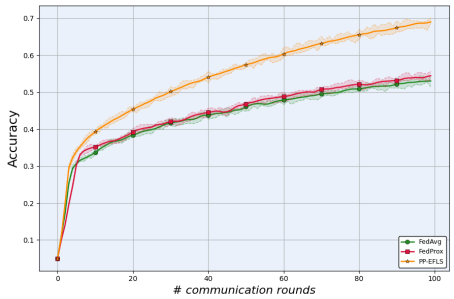


Figure 15: PoC; $\alpha = 10000$; LS=5

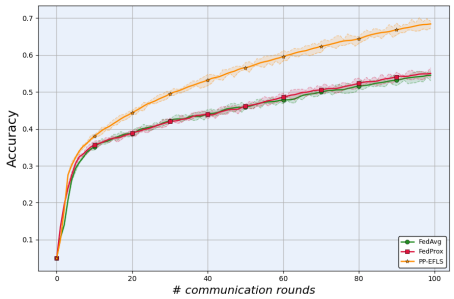


Figure 16: FedCBS; $\alpha = 10000$; LS=5

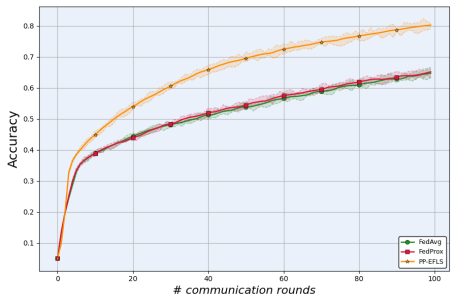


Figure 17: PoC; $\alpha = 10000$; LS=10

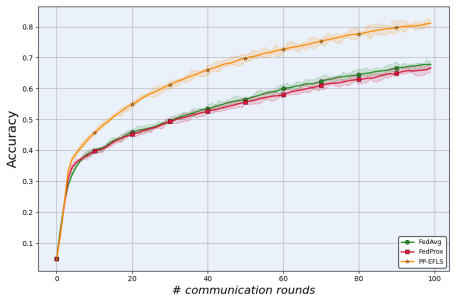


Figure 18: FedCBS; $\alpha = 10000$; LS=10

Discussion. *PP-EFLS* consistently reaches target accuracy in fewer epochs than *FedAvg* and *FedProx* across all settings, with the largest gains under severe heterogeneity ($\alpha = 0.1$), where the baselines often fail to converge within 100 epochs under *PoC* sampling while *PP-EFLS* does not.

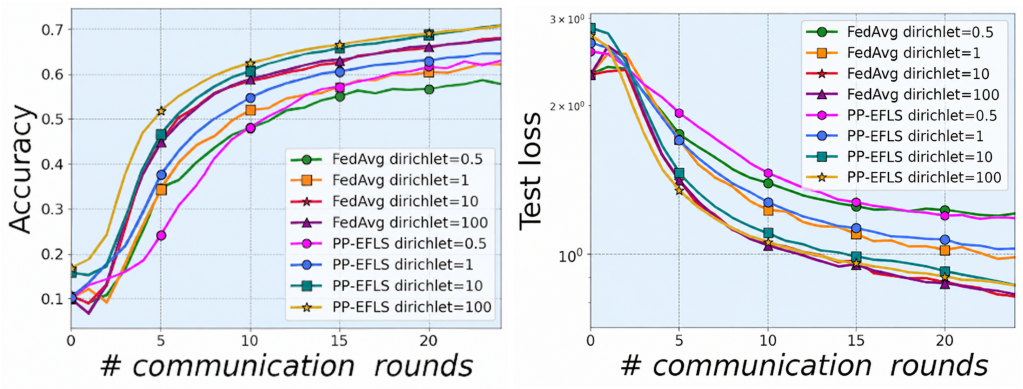
This gap narrows as data becomes more homogeneous, since all methods converge faster in absolute terms. *FedCBS* sampling substantially speeds up *FedAvg* and *FedProx* by mitigating participation bias, but has little effect on *PP-EFLS* – suggesting its gradient surrogate already provides much of this debiasing internally, without relying on a specific sampling strategy. Finally, more local steps reduce epoch counts for all methods, and PP-EFLS's advantage holds consistently across this dimension as well.

D.2 Experiment results with large-scale system

In addition to experiments with a small number of clients, we evaluated the proposed method in a larger-scale federated setting involving 120 clients. This setting more closely reflects practical federated learning deployments, where a central server typically coordinates training across a large and heterogeneous population of devices.

Algorithm	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
PP-EFLS	0.63	0.65	0.70	0.71
FedAvg	0.58	0.62	0.68	0.69

Table 18: Best accuracy over communication rounds



Discussion. In this larger-scale setting with 120 clients, *PP-EFLS* consistently outperforms *FedAvg* across all heterogeneity levels, confirming that the method's benefits persist beyond the small-scale experiments. As expected, accuracy improves for both algorithms as α increases (i.e., as data becomes more homogeneous), but *PP-EFLS*'s advantage over *FedAvg* is largest under strong heterogeneity ($\alpha = 0.5$) and gradually narrows as α grows, mirroring the trend observed in the smaller-scale experiments. This suggests that the method's ability to handle biased client participation scales well to larger, more realistic federated populations rather than being an artifact of the small-client setting.

D.3 Setup

D.3.1 Data distributions

We model data heterogeneity across clients using a Dirichlet distribution parameterized by concentration coefficient α, as it provides a principled and widely adopted mechanism for controlling label-skew. In this construction, each client's class proportions are sampled from a Dirichlet distribution, where smaller values of α induce highly sparse and imbalanced label distributions, thereby representing strong statistical heterogeneity across clients. Conversely, larger values of α yield increasingly uniform class proportions, approximating the IID setting as $\alpha \to \infty$. Specifically, we consider $\alpha = 0.1, 5, 10000$ to span a broad spectrum of heterogeneity regimes: $\alpha = 0.1$ corresponds to highly non-IID conditions with severe label skew and limited class overlap between clients, $\alpha = 5$ represents moderate heterogeneity with partially shared class distributions, and $\alpha = 10000$ approximates an almost IID scenario where client data distributions are effectively homogeneous.

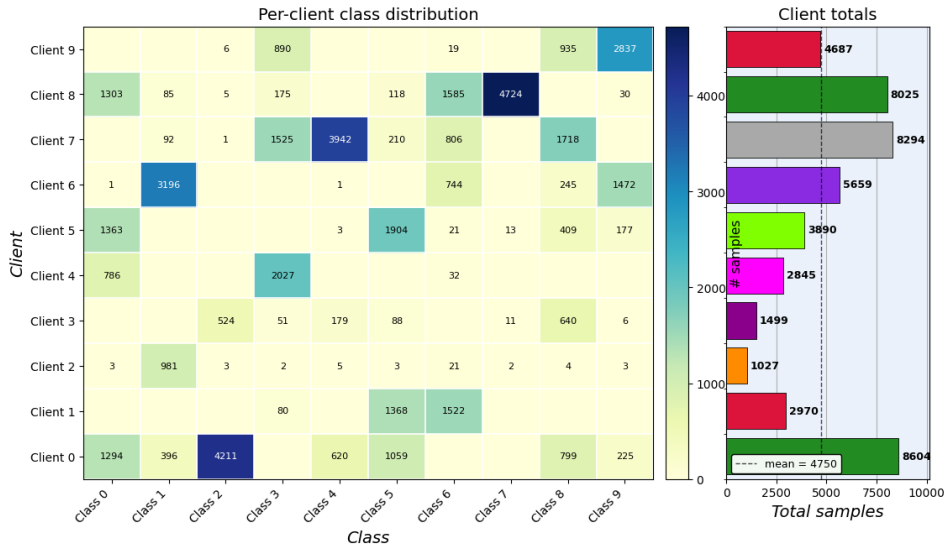


Figure 19: Client distribution over 10 clients with $\alpha = 0.1$.



Figure 20: Client distribution over 10 clients with $\alpha = 5$.

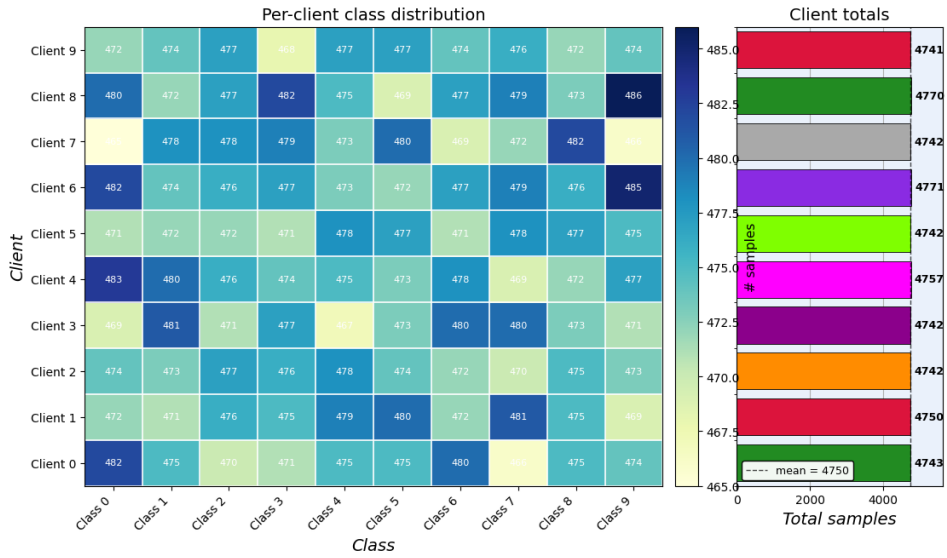


Figure 21: Client distribution over 10 clients with $\alpha = 10000$.

D.3.2 RESOURCES

All experiments were conducted on a dedicated computing cluster equipped with 8 GPUs, specifically NVIDIA A100. Depending on availability, each experiment was assigned a fixed number of GPU resources to maintain consistent computational conditions across all runs. To ensure fairness across experimental settings, all methods are evaluated under identical hardware configurations, and no approach is granted additional computational resources or asynchronous advantages.