A Unified Analysis of Federated Learning with Arbitrary Client Participation

Shiqiang Wang
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
wangshiq@us.ibm.com

Mingyue Ji
Department of ECE, University of Utah
Salt Lake City, UT 84112
mingyue.ji@utah.edu

Abstract

Federated learning (FL) faces challenges of intermittent client availability and computation/communication efficiency. As a result, only a small subset of clients can participate in FL at a given time. It is important to understand how partial client participation affects convergence, but most existing works have either considered idealized participation patterns or obtained results with non-zero optimality error for generic patterns. In this paper, we provide a unified convergence analysis for FL with arbitrary client participation. We first introduce a generalized version of federated averaging (FedAvg) that amplifies parameter updates at an interval of multiple FL rounds. Then, we present a novel analysis that captures the effect of client participation in a single term. By analyzing this term, we obtain convergence upper bounds for a wide range of participation patterns, including both non-stochastic and stochastic cases, which match either the lower bound of stochastic gradient descent (SGD) or the state-of-the-art results in specific settings. We also discuss various insights, recommendations, and experimental results.

1 Introduction

We consider a federated learning (FL) problem with N clients [17, 20, 38]. Our goal is to minimize:

$$f(\mathbf{x}) := \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^m$ is an m-dimensional model parameter. The local objective $F_n(\mathbf{x})$ of each client n is usually defined as $F_n(\mathbf{x}) := \mathbb{E}_{\xi_n \sim \mathcal{D}_n}\left[\ell_n(\mathbf{x}, \xi_n)\right]$, where $\mathcal{D}_n$ is the distribution of client n's local data, which cannot be observed globally because the data remains private, and $\ell_n(\mathbf{x}, \xi)$ is the per-sample loss function for parameter $\mathbf{x}$ and input data ξ. The objective (1) can be extended to a weighted average, but we do not write out the weights and consider them as part of $\ell_n(\mathbf{x}, \xi)$ and $F_n(\mathbf{x})$. A canonical way of solving (1) is to use federated averaging (FedAvg) [24], which is a form of stochastic gradient descent (SGD) that operates in multiple rounds, where each *round* includes multiple local update steps followed by communication between clients and a server to synchronize the updates.

Partial Participation. A major challenge in FL is that clients only intermittently participate in the collaborative training process [3]. The reason is twofold. First, FL applications usually have a large pool of clients. It is impractical to have all clients participating in all FL rounds, because it would require computation and communication that can consume a large amount of energy and also lead to network congestion. Second, clients may become unavailable from time to time. For example, if the client devices are smartphones, they may be willing to participate in FL only when they are charging (usually at night), to avoid draining the battery in outdoor environments where no charger is available.

Open Problem. The convergence of FL algorithms with realistic client participation patterns has not been fully understood. Most existing works on partial participation have only considered idealized scenarios, such as those where clients are randomly selected according to a given probability

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Summary and comparison of results (only showing baselines without constant error term)

Method	Participation	Convergence error upper bound	Remark
Yan et al. [35]	Unavailable $\leq E$ rounds	$\mathcal{O}\left(\frac{(1+\sigma^2)N^{1/4}\sqrt{E}}{\sqrt{ST}}\right)$	*Add'l assumption:* bounded gradient norm
Gu et al. [12]	Unavailable $\leq E$ rounds	$\mathcal{O}\left(\frac{(1+\sigma^2)\sqrt{E}}{\sqrt{NIT}}\right)$	*Add'l assumptions:* Hessian Lipschitz, a.s. bounded noise
Ours (this paper)	Regularized	$\mathcal{O}\left(\frac{\sigma}{\sqrt{SIT}}\right)$	Cor. 4.3, matching centralized SGD lower bound
	Ergodic	Approaches zero as $T \to \infty$	Prop. 4.4
	Mixing	$\mathcal{O}\left(\frac{1+\sigma^2}{\sqrt{SIT}}\right)$ i.E.; $\mathcal{O}\left(\frac{1+\sigma^2}{c\sqrt{SIT}}\right)$ w.p. $1-c$	Prop. 4.6 & Cor. 4.8, matching idealized bound
	Independent	$\mathcal{O}\left(\frac{1+\sigma^2}{\sqrt{SIT}}\right)$ i.E.; $\mathcal{O}\left(\frac{(1+\sigma^2)\ln(2/c)}{\sqrt{SIT}}\right)$ w.p. $1-c$	Prop. 4.7 & Cor. 4.8, matching idealized bound

Note: *We let $\rho = \frac{1}{\sqrt{S}}$ in our results for a direct comparison with other works*, see Section 5 for more details. We consider sufficiently large T in all cases and only show the dominant term in this table, with other variables ignored in $\mathcal{O}(\cdot)$. Definitions: i.E. = in expectation, w.p. = with probability, a.s. = almost surely, S: number of participating clients in each round ($S \leq N$), other definitions are given throughout the paper.

distribution that is independent across rounds [4, 8, 9, 18, 21, 22, 36]. They cannot incorporate cases where clients are unavailable for a period of time. Some recent results considering general scenarios either include a constant error term that does not vanish at convergence [5, 37] or are otherwise less competitive than state-of-the-art results obtained in idealized settings [12, 35]. Therefore, we ask:
1. What classes of participation patterns have guaranteed convergence to zero error?
2. For these classes, can the convergence rate match with that obtained in idealized scenarios?

To answer these questions, we need to overcome two challenges. First, we need to devise a *unified analytical methodology* for obtaining the convergence rates under various participation patterns. Second, we need a *general and realistic algorithm* for cross-device FL, with appropriate control options that can be configured to make the convergence rate competitive.

Our Contributions. In this paper, we overcome these challenges and give answers to the two questions above. We first introduce a generalized version of FedAvg, which *amplifies* the parameter updates after every P rounds, for some $P \geq 1$. When $P = 1$, our setting is the same as FedAvg with two-sided learning rates [18, 36], but we focus on an extended setup where P can be greater than one. For this generalized FedAvg algorithm, we perform a new convergence analysis that *unifies* the effect of partial participation into a single term ($\tilde{\delta}^2(P)$ in Section 3). This allows us to *decouple* our main convergence analysis from the analysis on partial participation, which largely simplifies the analytical procedure. Our novel methodology gives the results shown in Table 1, for a broad range of participation patterns. These patterns include *non-stochastic and regularized* participation, where all clients participate equally but not simultaneously within P rounds, and *stochastic* participation settings including *ergodic*, *stationary and strongly mixing* (e.g., Markov process), and *independent*. For the stochastic participation, we provide both expected and high probability convergence bounds, where the expected rates match the state-of-the-art FedAvg convergence rates that were derived for the idealized setting of random participation following an independent distribution. Furthermore, we provide new insights related to non-IID data, amplification of updates, "linear speedup" [39], etc.

To summarize, our main contributions are as follows:

- We introduce a generalized FedAvg algorithm which amplifies the updates aggregated over multiple rounds, where the amplification interval P can be tuned for the best convergence.
- We present novel analysis and unified methodology for obtaining competitive convergence upper bounds with arbitrary client participation patterns.
- We discuss important insights from both the theoretical and experimental results.

Related Work. FedAvg [24] is characterized by partial client participation and multiple local updates in each round. In the case of full participation, FedAvg is also known as local SGD [10, 14, 23, 28–30] and parallel restarted SGD [40]. Partial participation was considered in [4, 8, 9, 18, 21, 22, 36], with the assumption that clients are selected to participate probabilistically according to a given distribution that is independent across rounds. Some recent works have started to incorporate unavailable clients.

For example, the work in [27] allows inactive clients, but the theoretical result does not guarantee convergence if a client participates with zero probability in a certain round. Similarly, the result in [5, 37] includes a constant error term in the case of general participation in cross-device FL. Other works have obtained convergence rates related to the maximum number of inactive rounds, based on additional assumptions such as Hessian Lipschitz [12] and bounded gradient norm [35]. Their theory also requires all clients to participate once at the beginning (see [12, Remark 5.2]). In contrast, our result in this paper is based on a minimal set of assumptions. We consider a simple algorithm without requiring initial full participation, and we provide a better rate of convergence to zero error (see Table 1). Moreover, existing methods usually have separate analyses and results for each specific class of participation pattern, whereas we present a unified framework that largely simplifies the analytical procedure. Some more discussions on related works are in Appendix A.

2 Generalized FedAvg with Amplified Updates

We consider a generalized version of FedAvg as shown in Algorithm 1. In this algorithm, we define q_t^n to be the *participation weight* of client n in round t. If $q_t^n = 0$, the client does not participate in this round. Each client n participates whenever $q_t^n > 0$, where the weight q_t^n is applied in the global update step in Line 9. For mathematical convenience in our analysis, we assume in Algorithm 1 that all clients compute their local updates in Lines 4–8, regardless of whether they participate in the current round or not. We emphasize that this is *logically equivalent* to the practical setting where non-participating clients do not compute, because their computations have no effect in subsequent rounds when $q_t^n = 0$ (see Line 9). In every round t, each client n computes I steps of local updates according to Line 7, where $\gamma > 0$ is the local learning rate and $\mathbf{g}_n(\mathbf{y}_{t,i}^n)$ is the stochastic gradient of $F_n(\mathbf{y}_{t,i}^n)$ such that $\mathbb{E}\left[\mathbf{g}_n(\mathbf{y}_{t,i}^n) \middle| \mathbf{y}_{t,i}^n\right] = \nabla F_n(\mathbf{y}_{t,i}^n)$.

Starting at round t_0, the global updates are accumulated in $\mathbf{u}$ for P rounds (Line 10). At the end of every interval of P rounds, an amplification is applied to the accumulated updates (Line 12). Since $\mathbf{x}_t$ is first updated without amplification in rounds $t_0, \ldots, t_0 + P - 1$, Line 12 adds $\mathbf{u}$ after multiplying $\eta - 1$, which is equivalent to setting $\mathbf{x}_{t+1}$ to $\mathbf{x}_{t_0} + \eta\mathbf{u}$. When $\eta = 1$, there is no amplification and Line 12 has no effect in this case. In general, we allow any $\eta > 0$ including values of η that are less than one. When $\eta < 1$, we reduce (instead of amplify) the updates. Intuitively, amplification with $\eta > 1$ is usually preferred over reduction, because we can compute the gradients on similar model parameters in each round where only a small number of clients participates. After P rounds, the accumulated updates provide a better estimation for the overall client population, and amplifying the updates allows the model parameter to progress towards the descent direction for the majority of clients (see Appendix B.1 for an example). Also note that although P is a parameter in our algorithm, the dominant terms of our convergence upper bounds shown in Table 1 *do not* depend on P.

Algorithm 1: Generalized FedAvg with amplified updates and arbitrary participation

1 **Input:** $\gamma, \eta, \mathbf{x}_0, I, P, T$; **Output:** $\{\mathbf{x}_t : \forall t\}$
2 Initialize $t_0 \leftarrow 0$, $\mathbf{u} \leftarrow \mathbf{0}$;
3 **for** $t = 0, \ldots, T-1$ **do**
4 **for** $n = 1, \ldots, N$ *in parallel* **do**
5 $\mathbf{y}_{t,0}^n \leftarrow \mathbf{x}_t$;
6 **for** $i = 0, \ldots, I-1$ **do**
7 $\mathbf{y}_{t,i+1}^n \leftarrow \mathbf{y}_{t,i}^n - \gamma \mathbf{g}_n(\mathbf{y}_{t,i}^n)$
8 $\Delta_t^n \leftarrow \mathbf{y}_{t,I}^n - \mathbf{x}_t$;
9 $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \sum_{n=1}^N q_t^n \Delta_t^n$; //update
10 $\mathbf{u} \leftarrow \mathbf{u} + \sum_{n=1}^N q_t^n \Delta_t^n$; //accumulate
11 **if** $t + 1 - t_0 = P$ **then**
12 $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1} + (\eta - 1)\mathbf{u}$; //amplify
13 $t_0 \leftarrow t + 1$;
14 $\mathbf{u} \leftarrow \mathbf{0}$;

3 Convergence Analysis and Main Result

We analyze Algorithm 1 as follows. For ease of analysis, we assume that $\sum_{n=1}^N q_t^n = 1$ for $q_t^n \geq 0$ for all t.[1] We also define ρ such that $\sum_{n=1}^N (q_t^n)^2 \leq \rho^2$ for all t. As $\sum_{n=1}^N q_t^n = 1$, there always

[1] When $\sum_{n=1}^N q_t^n \neq 1$, the algorithm is equivalent to the case of $\{q_t^n\}$ normalized to one and a global learning rate applied to the updates in each round. Since we already amplify the updates every P rounds, we do not apply a separate global learning rate to each round.

exists ρ such that $\rho^2 \leq 1$. We define $\mathcal{Q} := \{q_t^n : \forall n, t\}$. Mathematically, we use the conditional expectation $\mathbb{E}\left[\cdot | \mathcal{Q}\right]$ to denote the case where $\mathcal{Q}$ is given and the expectation is only over the stochastic gradient noise. In the case where $\mathcal{Q}$ is stochastic, the full expectation $\mathbb{E}\left[\cdot\right]$ is taken over both the noise and $\mathcal{Q}$. We make the following minimal set of assumptions that are commonly used in the literature.

Assumption 1 (Lipschitz gradient)**.**

$$\left\|\nabla F_n(\mathbf{x}) - \nabla F_n(\mathbf{y})\right\| \leq L \left\|\mathbf{x} - \mathbf{y}\right\|, \forall \mathbf{x}, \mathbf{y}, n. \tag{2}$$

Assumption 2 (Unbiased stochastic gradient with bounded variance)**.**

$$\mathbb{E}\left[\mathbf{g}_n(\mathbf{x}) | \mathbf{x}\right] = \nabla F_n(\mathbf{x}) \text{ and } \mathbb{E}\left[\left\|\mathbf{g}_n(\mathbf{x}) - \nabla F_n(\mathbf{x})\right\|^2 \middle| \mathbf{x}\right] \leq \sigma^2, \forall \mathbf{x}, n. \tag{3}$$

Assumption 3 (Bounded gradient divergence)**.**

$$\left\|\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq d^2, \forall \mathbf{x}, n. \tag{4}$$

The gradient divergence in Assumption 3 is related to the degree of non-IID data distribution across clients. To better interpret how different divergence components affect the convergence, we introduce an alternative set of divergence bounds in Assumption 3' as follows. We will show later (in Section 4) that if Assumption 3 holds, then Assumption 3' also holds with properly chosen $\tilde{\beta}^2$, $\tilde{\nu}^2$, and $\tilde{\delta}^2(P)$.

Assumption 3' (Alternative gradient divergence bound)**.**

$$\left\|\sum_{n=1}^N q_t^n \left[\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\right]\right\|^2 \leq \tilde{\beta}^2, \forall \mathbf{x}, t, \tag{5}$$

$$\sum_{n=1}^N q_t^n \left\|\nabla F_n(\mathbf{x}) - \sum_{n'=1}^N q_t^{n'} \nabla F_{n'}(\mathbf{x})\right\|^2 \leq \tilde{\nu}^2, \forall \mathbf{x}, t, \tag{6}$$

$$\left\|\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} \sum_{n=1}^N q_t^n \left(\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\right)\right\|^2 \leq \tilde{\delta}^2(P), \forall \mathbf{x}, t_0. \tag{7}$$

In the above, $\tilde{\beta}^2$ and $\tilde{\nu}^2$ capture the gradient divergence in an arbitrary round t, and $\tilde{\delta}^2(P)$ captures the divergence of time-averaged gradients over P rounds, which is a function of P. In particular, $\tilde{\beta}^2$ is an upper bound of the divergence between the original objective's gradient $\nabla f(\mathbf{x})$ and the averaged gradient among participating clients weighted by $\{q_n^t\}$. When all the clients participate, we have $q_t^n = \frac{1}{N}$ and (5) holds with $\tilde{\beta}^2 = 0$. The quantity $\tilde{\nu}^2$ is an upper bound of the divergence between each client's gradient and the averaged gradient of participating clients, where the average is also weighted by $\{q_n^t\}$. We will show in Section 4.1 that the overall divergence d^2 in Assumption 3 can be expressed as a sum of $\tilde{\beta}^2$ and $\tilde{\nu}^2$, as intuition suggests. The quantity $\tilde{\delta}^2(P)$ in (7) extends $\tilde{\beta}^2$ in (5) by computing the average over P rounds (inside the norm) instead of a single round. When the weights $\{q_n^t\}$ are properly chosen, as P gets large, the average q_n^t over P rounds gets close to each other for different n, and $\tilde{\delta}^2(P)$ becomes small. We will formally analyze this behavior in Section 4.2.

In the following, let $\mathcal{F} := f(\mathbf{x}_0) - f^*$, where f^* is the true minimum of (1), i.e., $f^* := \min_{\mathbf{x}} f(\mathbf{x})$.

Theorem 3.1. *When Assumptions 1, 2, and 3' hold, $\gamma \leq \frac{1}{12LIP}$, $\gamma\eta \leq \frac{1}{LIP}$, and $P \leq \frac{T}{2}$, we have*

$$\min_t \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{\mathcal{F}}{\gamma\eta IT} + \gamma^2 L^2 I^2 \tilde{\nu}^2 + \gamma^2 L^2 I^2 P^2 \tilde{\beta}^2 + \tilde{\delta}^2(P) + \left(\gamma^2 L^2 I + \gamma^2 L^2 I P \rho^2 + \gamma\eta L \rho^2\right)\sigma^2\right). \tag{8}$$

Proof Sketch. We first use smoothness to relate $f(\mathbf{x}_{t_0+P})$ and $f(\mathbf{x}_{t_0})$, which includes an inner-product term that can be expanded into several terms. One of these terms includes $\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla f(\mathbf{x}_{t_0})$. We upper bound this term with three other terms including $\left\|\mathbf{y}_{t,i}^n - \sum_{n'=1}^N q_t^{n'} \mathbf{y}_{t,i}^{n'}\right\|^2$, $\left\|\sum_{n'=1}^N q_t^{n'} \mathbf{y}_{t,i}^{n'} - \mathbf{x}_{t_0}\right\|^2$, and $\tilde{\delta}^2(P)$. The upper bounds of the first two terms are found by solving recurrence relations, where the recurrence for the second term includes the first term. This nested recurrence is a uniqueness in this proof. The full proof is given in Appendix C. □

For two specific (but different) learning rate configurations, we can obtain the following corollaries.

Corollary 3.2. *Choosing* $\gamma = \frac{1}{12LIP\sqrt{T}}$ *and* $\eta = \min\left\{\frac{12P\sqrt{LI\mathcal{F}}}{\sigma\rho}; 12\sqrt{T}\right\}$*, for* $P \leq \frac{T}{2}$*, we have*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{\sigma\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{LP\mathcal{F}}{T} + \frac{\tilde{\nu}^2}{P^2T} + \frac{\tilde{\beta}^2}{T} + \tilde{\delta}^2(P) + \frac{\sigma^2}{IPT}\right). \tag{9}$$

Corollary 3.3. *If* $\frac{\sqrt{\mathcal{F}}}{\rho\sqrt{LIT}} \leq \frac{1}{LIP}$*, choosing* $\gamma = \frac{1}{12LIP\sqrt{T}}$ *and* $\eta = \frac{12P\sqrt{LI\mathcal{F}}}{\rho}$*, for* $P \leq \frac{T}{2}$*, we have*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{(1+\sigma^2)\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{\tilde{\nu}^2}{P^2T} + \frac{\tilde{\beta}^2}{T} + \tilde{\delta}^2(P) + \frac{\sigma^2}{IPT}\right). \tag{10}$$

These two corollaries are key building blocks of our *unified convergence analysis framework*. The choice between them depends on whether and how fast P grows with T. We will also see in the next section that $\tilde{\delta}^2(P)$ is key in capturing the effect of different participation patterns. After deriving $\tilde{\delta}^2(P)$ such that (7) holds, plugging it back to one of the above corollaries gives the results in Table 1.

4 Interpreting and Applying the Unified Framework

4.1 Discussion on $\tilde{\beta}^2$ and $\tilde{\nu}^2$: Decomposition of Divergence

In Corollaries 3.2 and 3.3, we have an interesting observation that the term involving $\tilde{\nu}$ decreases with P, whereas the term involving $\tilde{\beta}$ does not decrease with P. This suggests that when P is large, the divergence term $\tilde{\beta}$ has the main effect on convergence. We recall that $\tilde{\nu}$ captures the divergence between the gradient of each client and the average gradient of clients that participate in a round t. Since the learning rate γ is inversely proportional to both I and P, a large P gives a smaller learning rate which makes the divergence (related to $\tilde{\nu}$) after I iterations in each round less significant. For the term with $\tilde{\beta}$, increasing P does not have the same effect, because parameter averaging is only conducted across the participating clients instead of all the clients, and $\tilde{\beta}$ captures the divergence with respect to all the clients. The effect of all clients' participation pattern is captured by the term $\tilde{\delta}^2(P)$.

From Assumption 3, by adding and subtracting $\sum_{n'=1}^N q_t^{n'} \nabla F_{n'}(\mathbf{x})$ inside the norm and expanding the square, we can obtain

$$\begin{aligned}&\textstyle\sum_{n=1}^N q_t^n \|\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \\ &= \textstyle\sum_{n=1}^N q_t^n \left\|\nabla F_n(\mathbf{x}) - \sum_{n'=1}^N q_t^{n'} \nabla F_{n'}(\mathbf{x})\right\|^2 + \left\|\sum_{n'=1}^N q_t^{n'} \nabla F_{n'}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq d^2,\end{aligned} \tag{11}$$

because $\sum_{n=1}^N q_t^n = 1$ and the inner-product term is zero. The two terms in the right-hand side (RHS) of (11) are equal to the left-hand side (LHS) of (5) and (6), respectively. We thus have the following result showing that $\tilde{\beta}^2$ and $\tilde{\nu}^2$ can be seen as a decomposition of the original divergence d^2.

Proposition 4.1. *There exist* $\tilde{\beta}^2$ *and* $\tilde{\nu}^2$*, such that* $\tilde{\beta}^2 + \tilde{\nu}^2 \leq d^2$ *while satisfying* (5) *and* (6).

4.2 Discussion on $\tilde{\delta}^2(P)$: Effect of Partial Participation

The results in Theorem 3.1 and Corollaries 3.2 and 3.3 include a term of $\tilde{\delta}^2(P)$. To guarantee convergence to zero error, $\tilde{\delta}^2(P)$ needs to be either equal to zero or decrease in T.

Condition for $\tilde{\delta}^2(P) = 0$. From (1), we know that $0 = \frac{1}{N}\sum_{n=1}^N F_n(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^N (F_n(\mathbf{x}) - f(\mathbf{x}))$. Hence,

$$\mu_{t_0} \textstyle\sum_{n=1}^N (F_n(\mathbf{x}) - f(\mathbf{x})) = 0 \tag{12}$$

for an arbitrary constant μ_{t_0}, which can possibly depend on t_0. Then, we obtain

$$\begin{aligned}\left\|\tfrac{1}{P}\textstyle\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^N q_t^n (\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|^2 &= \left\|\textstyle\sum_{n=1}^N \frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n (\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|^2 \\ &= \left\|\textstyle\sum_{n=1}^N \left[\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n - \mu_{t_0}\right](\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|^2,\end{aligned} \tag{13}$$

where the last equality is due to (12). Since (13) is equal to the LHS of (7), we have the following.

Proposition 4.2 (Regularized participation)**.** *If* $\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n - \mu_{t_0} = 0$ *for some* μ_{t_0} *and all* $n \in \{1,\dots,N\}$, $t_0 \in \{0, P, 2P, \dots\}$, *then* (7) *holds with* $\tilde{\delta}^2(P) = 0$.

This implies that if P is chosen so that the averaged participation weights over every interval of P rounds ($\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n$) are equal to each other among all the clients $n \in \{1,\dots,N\}$, i.e., regularized, then $\tilde{\delta}^2(P) = 0$. Together with Corollary 3.2 and Proposition 4.1, we have the following.

Corollary 4.3 (Regularized participation)**.** *If the condition of Proposition 4.2 holds for* $P \leq \frac{T}{2}$, *and we choose the learning rates according to Corollary 3.2, then*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{\sigma\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{LP\mathcal{F} + d^2 + \sigma^2}{T}\right). \tag{14}$$

If Proposition 4.2 holds for a finite P that does not depend on T, we obtain a convergence rate of $\mathcal{O}\left(\frac{\sigma\rho}{\sqrt{IT}} + \frac{1+\sigma^2}{T}\right)$ for $T \geq 2P$, where all the other constants are ignored in the $\mathcal{O}(\cdot)$ notation here. For full gradient descent with $\sigma^2 = 0$, this convergence rate is improved to $\mathcal{O}\left(\frac{1}{T}\right)$.

Interpreting $\tilde{\delta}^2(P)$ ***Using Variance.*** The condition for $\tilde{\delta}^2(P) = 0$ may be too stringent for some practical scenarios of FL. Next, we show that $\tilde{\delta}^2(P)$ can be expressed as the variance of client participation weights averaged over P rounds. We first express $\tilde{\delta}^2(P)$ with d^2 by further bounding:

$$(13) \leq N\sum_{n=1}^{N}\left[\overline{q_{t_0}^n} - \mu_{t_0}\right]^2 d^2, \tag{15}$$

where we first use Jensen's inequality on the sum over n in (13), then move the term $\overline{q_{t_0}^n} - \mu_{t_0}$ to the outside of the norm, where we define $\overline{q_{t_0}^n} := \frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n$ for any $n \in \{1,\dots,N\}$, and afterwards bound $\|\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\|^2$ by d^2 according to Assumption 3. We note that (7) holds if the RHS of (15) is upper bounded by $\tilde{\delta}^2(P)$. Therefore, we can obtain $\tilde{\delta}^2(P)$ by analyzing the statistical properties of $\overline{q_{t_0}^n} - \mu_{t_0}$ and then choosing $\tilde{\delta}^2(P)$ to be equal to the RHS of (15).

In the following, we assume that μ_{t_0} is chosen such that $\mu_{t_0} = \mathbb{E}\left[q_t^n\right]$ for all $n \in \{1,\dots,N\}$ and $t \in \{t_0, t_0+1, \dots, t_0+P-1\}$. This implies that the mean participation weights of all the clients within the same cycle of P rounds are equal. Note that this condition is the same as unbiased client sampling in existing works [8, 9, 18, 21, 22, 36], because we consider an unweighted average in (1) and any weighting is included in the local objective function $F_n(\mathbf{x})$. However, differently from these existing works, we do not assume independence here (we will only assume independence in a special case later). Under this condition, it is apparent that $\mathrm{Var}\left(\overline{q_{t_0}^n}\right) = \mathbb{E}\left[\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2\right]$ is the variance of q_t^n averaged over P rounds, for any $n \in \{1,\dots,N\}$. We immediately have the following result.

Proposition 4.4 (Ergodic participation)**.** *If* $\{q_t^n : \forall t\}$ *is a mean-ergodic process for any* $n \in \{1,\dots,N\}$, *which means that* $\lim_{P\to\infty}\mathbb{E}\left[\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2\right] = 0$ *for any* n *and* t_0, *then there exists* $\tilde{\delta}^2(P)$ *such that* $\lim_{P\to\infty}\tilde{\delta}^2(P) = 0$ *while satisfying* (7) *in expectation.*

Next, we consider the class of stationary and strongly mixing processes. Informally, a random process is said to be *strongly mixing* with coefficient $\alpha(P)$ if the outcomes that are at least P steps apart are nearly independent, where the distance between the joint probability distribution of the outcomes and their independent counterpart is at most $\alpha(P)$. See [2, Sec. 27] for a formal definition. A specific example of stationary and strongly mixing processes is finite-state irreducible and aperiodic Markov chains, which have an $\alpha(P)$ that exponentially decreases in P [2, Thm. 8.9]. It is known that the following (generalized) central limit theorem holds for stationary and strongly mixing processes, where we adapt the result to our participation weights.

Lemma 4.5 ([2, Thm. 27.4])**.** *If* $\{q_t^n : \forall t\}$ *is stationary and strongly mixing with* $\alpha(P) = \mathcal{O}(P^{-5})$, *for any* $n \in \{1,\dots,N\}$, *then* $\overline{q_{t_0}^n} \sim \mathcal{N}\left(\mu_{t_0}, \frac{\hat{\upsilon}^2}{P}\right)$ *for any* t_0 *when* $P \to \infty$, *where* $\mathcal{N}(\cdot,\cdot)$ *denotes the normal distribution and* $\hat{\upsilon}^2 := \mathrm{Var}(q_{t_0}^n) + 2\sum_{p=1}^{\infty}\mathrm{Cov}(q_{t_0}^n, q_{t_0+p}^n)$.

From the definition of strongly mixing, we know that $\mathrm{Cov}(q_{t_0}^n, q_{t_0+p}^n)$ in the definition of $\hat{\upsilon}^2$ approaches zero as p gets large, hence $\sum_{p=1}^{\infty}\mathrm{Cov}(q_{t_0}^n, q_{t_0+p}^n)$ converges to a finite value. When $P \to \infty$

and $\alpha(P) = \mathcal{O}(P^{-5})$, Lemma 4.5 shows that $\mathrm{Var}\left(\overline{q_{t_0}^n}\right) = \frac{\hat{\upsilon}^2}{P}$. Chebyshev's inequality gives

$$\Pr\left\{\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2 \leq \frac{\hat{\upsilon}^2}{cP}\right\} \geq 1 - c. \tag{16}$$

Plugging $\mathrm{Var}\left(\overline{q_{t_0}^n}\right) = \mathbb{E}\left[\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2\right] = \frac{\hat{\upsilon}^2}{P}$ and $\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2 \leq \frac{\hat{\upsilon}^2}{cP}$ into the RHS of (15) gives the following expected and high-probability results, respectively, where "w.p." stands for "with probability".

Proposition 4.6 (Stationary and strongly mixing participation)**.** *If $\{q_t^n : \forall t\}$ is stationary and strongly mixing with $\alpha(P) = \mathcal{O}(P^{-5})$, for any $n \in \{1, \ldots, N\}$, then as $P \to \infty$:*
Choosing $\tilde{\delta}^2(P) = \frac{N^2 d^2 \hat{\upsilon}^2}{P}$ satisfies (7) *in expectation; $\tilde{\delta}^2(P) = \frac{N^2 d^2 \hat{\upsilon}^2}{cP}$ satisfies* (7) *w.p. $1 - c$.*

For independent participation, we can obtain a similar result while *not* requiring $P \to \infty$. Assuming that $\mathrm{Var}(q_t^n) \leq \upsilon^2$ for any t and n, it is evident that $\mathrm{Var}\left(\overline{q_{t_0}^n}\right) \leq \frac{\upsilon^2}{P}$, because the variance of the sum of independent random variables is equal to the sum of the variance, and $\mathrm{Var}(aZ) = a^2\mathrm{Var}(Z)$, for an arbitrary constant a and random variable Z. We note that Hoeffding's inequality [16] gives

$$\Pr\left\{\left(\overline{q_{t_0}^n} - \mu_{t_0}\right)^2 \leq \frac{\ln(2/c)}{2P}\right\} \geq 1 - c. \tag{17}$$

We have the following result by plugging into the RHS of (15).

Proposition 4.7 (Independent participation)**.** *If $\{q_t^n : \forall t\}$ is independent across t, for any n, then:*
Choosing $\tilde{\delta}^2(P) = \frac{N^2 d^2 \upsilon^2}{P}$ satisfies (7) *in expectation; $\tilde{\delta}^2(P) = \frac{N^2 d^2 \ln(2/c)}{2P}$ satisfies* (7) *w.p. $1-c$.*

Note that the independence here is only assumed across t, so q_t^n and $q_t^{n'}$ for the same t but $n \neq n'$ can still be dependent of each other. Compared to Proposition 4.6, the high-probability bound in Proposition 4.7 has c in the logarithmic term, which is tighter since $\frac{1}{2}\ln\left(\frac{2}{c}\right) \leq \frac{1}{c} - \frac{1}{2}$ (ignoring $\hat{\upsilon}^2$).

Choosing P as a Function of T. The choices of $\tilde{\delta}^2(P)$ in Propositions 4.6 and 4.7 include P in the denominator. Hence, we can guarantee convergence to zero error if we choose P to be an increasing function of T, which also ensures that $P \to \infty$ as $T \to \infty$. If we choose $P \propto N^{\frac{5}{2}}\sqrt{IT}$, we can obtain the following convergence rate from Corollary 3.3 together with Proposition 4.1.

Corollary 4.8 (Stochastic participation)**.** *If the conditions of either Proposition 4.6 or Proposition 4.7 hold, and we choose $P = \hat{\upsilon}^2 N^{\frac{5}{2}}\sqrt{IT}$ or $P = \upsilon^2 N^{\frac{5}{2}}\sqrt{IT}$, respectively, and the learning rates according to Corollary 3.3, then the convergence error in expectation over $\mathcal{Q}$ satisfies*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{(1+\sigma^2)\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{d^2}{\sqrt{NIT}} + \frac{d^2+\sigma^2}{T}\right), \tag{18}$$

for $I \leq \frac{\rho}{\hat{\upsilon}^2 N^{5/2}\sqrt{L\mathcal{F}}}$ and $T \to \infty$ when using conditions of Proposition 4.6, or $I \leq \frac{\rho}{\upsilon^2 N^{5/2}\sqrt{L\mathcal{F}}}$ and $T \geq 4\upsilon^4 I N^5$ when using condition of Proposition 4.7.

Remark. Although the upper bound of I in Corollary 4.8 may appear restrictive, note that the optimal solution $\mathbf{x}^*$ to (1) remains the same when the objective $f(\mathbf{x})$ is scaled by a multiplicative positive constant, which in turn scales $\sqrt{L\mathcal{F}}$. More specifically, consider an arbitrary $a > 0$, we define $F'(\mathbf{x}) := aF(\mathbf{x})$ and $f'(\mathbf{x}) := \frac{1}{N}\sum_{n=1}^N F_n'(\mathbf{x}) = af(\mathbf{x})$. Then, we have $f'(\mathbf{x}_0) - f'^* = af(\mathbf{x}_0) - af^* = a\mathcal{F}$. We also have $\|\nabla F_n'(\mathbf{x}) - \nabla F_n'(\mathbf{y})\| = a\|\nabla F_n(\mathbf{x}) - \nabla F_n(\mathbf{y})\| \leq aL\|\mathbf{x} - \mathbf{y}\|$ from Assumption 1, due to the linearity of gradients. Hence, by choosing $a \in (0, 1)$ and replacing $F(\mathbf{x})$ with $F'(\mathbf{x})$, we can make $\sqrt{L\mathcal{F}}$ arbitrarily small and $\frac{\rho}{\hat{\upsilon}^2 N^{5/2}\sqrt{L\mathcal{F}}}$ or $\frac{\rho}{\upsilon^2 N^{5/2}\sqrt{L\mathcal{F}}}$ (the upper bound of I in Corollary 4.8) arbitrarily large without affecting the optimal solution $\mathbf{x}^*$. Thus, we can potentially allow arbitrarily large I by scaling $f(\mathbf{x})$ without changing the optimal solution $\mathbf{x}^*$. We leave the in-depth study of this phenomenon for future work, where we recognize that such a scaling will affect the LHS of the convergence bound too, although the optimal solution $\mathbf{x}^*$ does not change.

We also give the following further insight related to Corollary 4.8. In Corollary 4.8, we choose P to be proportional to the variance $\hat{\upsilon}^2$ or υ^2. This is intuitive because when the clients' participation weights have higher variance, we would like to wait for more rounds before amplifying, so that the contributions of clients are more balanced. In the same way, P increases with the number of clients N. Recall that $\sum_{n=1}^N q_t^n = 1$. Intuitively, when the participation patterns of individual clients remain

"unchanged", q_t^n scales with $\frac{1}{N}$ and its variance $\hat{\upsilon}^2$ or υ^2 scales with $\frac{1}{N^2}$, so the product $\hat{\upsilon}^2 N^{\frac{5}{2}} \sim \sqrt{N}$ or $\upsilon^2 N^{\frac{5}{2}} \sim \sqrt{N}$, implying that P effectively scales with $\sqrt{N}$ when this intuition holds.

While the result in Corollary 4.8 is in expectation, the high-probability bound is similar (see Table 1). In general, since (7) is used as an upper bound in the proof of Theorem 3.1, if $\tilde{\delta}^2(P)$ is chosen so that (7) holds in expectation or with a certain probability, such as in Propositions 4.4, 4.6 and 4.7, then the final convergence result also holds in expectation or with a certain probability, respectively.

5 Discussions and Insights

Interpretation of "Linear Speedup". Linear speedup is a desirable property seen in existing works that consider idealized client participation [36, 39, 40]. It essence, it means that the same convergence error can be achieved by increasing the number of participating clients S (with $S \leq N$) and reducing the number of rounds T, while keeping the product ST unchanged. In our case of arbitrary client participation, the coefficient ρ is a generalization of the $1/\sqrt{S}$ term for linear speedup in existing works that select a fixed number of S clients in each round. When S clients participate with equal weight, because $\sum_{n=1}^{N} q_t^n = 1$, we have $q_t^n = 1/S$ for the participating clients and $q_t^n = 0$ for the non-participating clients, so $\rho = \left[\sum_{n=1}^{N} (q_t^n)^2\right]^{1/2} = 1/\sqrt{S}$. Plugging this ρ back to the convergence results, we achieve a convergence rate of $\mathcal{O}\left(\sigma/\sqrt{SIT}\right)$ in Corollary 4.3 and $\mathcal{O}\left((1+\sigma^2)/\sqrt{SIT}\right)$ in Corollary 4.8, for sufficiently large T while ignoring the other variables in $\mathcal{O}(\cdot)$, as shown in Table 1. This shows that we can achieve linear speedup in S in this special case. We can also generalize to having at least $S_{\min}$ clients participating with equal weight in each round. Following the same argument, we have $\rho \leq 1/\sqrt{S_{\min}}$ in this case, giving a linear speedup in $S_{\min}$. In general, we achieve a linear speedup factor of $1/\rho^2$ in the case of arbitrary participation.

Matching Lower Bound or State-of-the-Art Results. We continue to assume sufficiently large T. For regularized participation (Corollary 4.3), our convergence rate of $\mathcal{O}\left(\sigma/\sqrt{SIT}\right)$ *matches the lower bound of convergence error for centralized SGD* [1]. The lower bound states that to reach a convergence error of ϵ, there exists an objective function such that at least $\Omega\left(\sigma^2/\epsilon^2\right)$ stochastic gradient oracle calls are required.[2] We have SIT oracle calls to reach $\epsilon = \mathcal{O}\left(\sigma/\sqrt{SIT}\right)$ according to our result, thus our upper bound matches this lower bound and is asymptotically optimal. Note that our upper bound still matches the lower bound if we include the coefficient $\sqrt{L\mathcal{F}}$ in $\mathcal{O}(\cdot)$ and $L\mathcal{F}$ in $\Omega(\cdot)$. For stochastic participation (Corollary 4.8), our convergence rate of $\mathcal{O}\left((1+\sigma^2)/\sqrt{SIT}\right)$ *matches state-of-the-art results of FedAvg* [8, 18, 36] that were obtained for the more idealized case of clients being selected to participate according to a specific independent sampling scheme.

Special Case of Waiting for All Clients. We consider a specific configuration of FedAvg where all the clients wait for P rounds before proceeding with SGD. In every iteration i of round t within these P rounds, each participating client n may call the stochastic gradient oracle to obtain $\mathbf{g}_n(\mathbf{x}_t)$, but it does not perform local updates. Instead, it accumulates and averages the sampled instances of $\mathbf{g}_n(\mathbf{x}_t)$. SGD is performed once after these P rounds using the *average* $\mathbf{g}_n(\mathbf{x}_t)$ from each client n. Assume that P is large enough so that each client n computes $\mathbf{g}_n(\mathbf{x}_t)$ at least $M \geq 1$ times that are averaged afterwards. If the M instances of $\mathbf{g}_n(\mathbf{x}_t)$ are all independent from each other, the noise σ^2 (defined in Assumption 2) becomes σ^2/M. This configuration can be considered as a special case of Algorithm 1, where all N clients perform SGD once in P rounds, so there are T/P updates in total, with a reduced stochastic gradient noise of σ^2/M. By replacing σ and SIT in our bound $\mathcal{O}\left(\sigma/\sqrt{SIT}\right)$ with $\sigma/\sqrt{M}$ and NT/P, respectively, we obtain a convergence rate of $\mathcal{O}\left(\sigma\sqrt{P}/\sqrt{MNT}\right)$ for regularized participation with this "wait-for-all" method.

In theory, when considering regularized participation and M is large enough, the convergence rate of this "wait-for-all" method can potentially match with that of the original configuration of Algorithm 1. In this way, the convergence error upper bound of both methods can match the lower bound of SGD convergence error (see discussion above). However, in practice, reducing the stochastic gradient noise as in this "wait-for-all" method may only give limited improvement. One reason is that practical settings have a finite number of training data samples, so the noise can only be reduced by a certain degree. In addition, having some noise in SGD can prevent the model parameter from being trapped in

[2] Note that [1] defines the convergence error as $\|\nabla f(\mathbf{x}_t)\|$ whereas we define it as $\|\nabla f(\mathbf{x}_t)\|^2$. Hence, the order of our ϵ is different from that in [1].

saddle points and local minima in practice. The empirical results in Section 6 show that Algorithm 1 with amplification performs better than waiting for all the clients. We also note that our results for stochastic participation (Corollary 4.8) *does not* require all the clients to participate within P rounds.

Achieving Regularized Participation in Practice. In theory, regularized participation requires that $\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} q_t^n$ is equal to each other for all the clients $n \in \{1, \ldots, N\}$, for a properly chosen P. When all the clients are connected to the system and client sampling is performed to limit the computation and communication overhead, this condition holds when the clients are selected according to a permutation, i.e., all the clients participate once before the same client can be selected again. In the case where clients get disconnected from time to time, the system can try to properly schedule the participation of connected clients and their weights $\{q_t^n\}$, so that within a cycle of P rounds, all the clients participate with equal averaged weights. In practice, it can be sufficient if this regularized condition is only approximately satisfied, especially when multiple clients have similar datasets so that $\tilde{\delta}^2(P)$ is small although not necessarily zero. The empirical results in Section 6 show that Algorithm 1 with amplification can give good performance even if P is less than the participation cycle of all clients. This suggests that, as long as the subset of clients that participate in P rounds are more representative of the overall data distribution than the (usually much smaller) subset of clients that participate in a single round, amplifying the updates every P rounds can be useful.

6 Experiments

We ran experiments of training convolutional neural networks (CNNs) with FashionMNIST [34] and CIFAR-10 [19] datasets, each of which has images in 10 different classes. We set the total number of clients to $N = 250$. Similar to existing works [6, 7, 36], we partition the data into clients so that each client has data of one majority class label, with 5% of data of other (minority) labels, to simulate a setup with non-IID data distribution which is often encountered in FL scenarios. We assume that the clients' availability exhibits a periodic pattern inspired by [6, 7]. Namely, in the first 100 rounds, only clients with the first two majority labels are available; in the next 100 rounds, only clients with the next two majority labels are available, and so on. In each round, $S = 10$ clients participate in a regularized manner, out of all the currently available clients. To speed up initial rounds of training, similar to practical deep learning implementations, we start with standard FedAvg with a relatively large initial learning rate. The initial rates are $\gamma = 0.1$ and $\gamma = 0.05$ without amplification (i.e., $\eta = 1$) for FashionMNIST and CIFAR-10, respectively, which were obtained using grid search in a separate scenario of always participation. After an initial training of $2,000$ rounds for FashionMNIST and $4,000$ rounds for CIFAR-10, we study the performance of different approaches with their own learning rates. Additional setup details and results are given in Appendix D.

Comparing Different Methods. We show the results of Algorithm 1 both with and without amplification. When using amplification, we set $\eta = 10$ and $P = 500$. We also compare to an algorithm that waits for all the clients, related to our discussion in Section 5, for two settings where each client computes its gradients either on a minibatch or on the entire dataset, which we refer to as "wait-minibatch" and "wait-full", respectively. The best learning rate γ of each approach was separately found on a grid that is $\{1, 0.1, 0.01, 0.001, 0.0001\}$ times the initial learning rate, with the periodic participation pattern described above. The results are shown in Figure 1.

We make a few key observations as follows. First, we clearly see that *Algorithm 1 with amplification gives the best performance*. By choosing $P = 500$, we match with the participation period of clients,

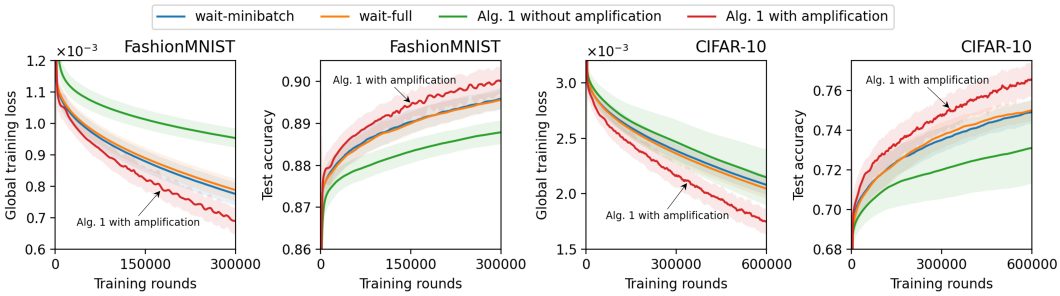


Figure 1: Results for different approaches with periodically connected clients ($P = 500$).

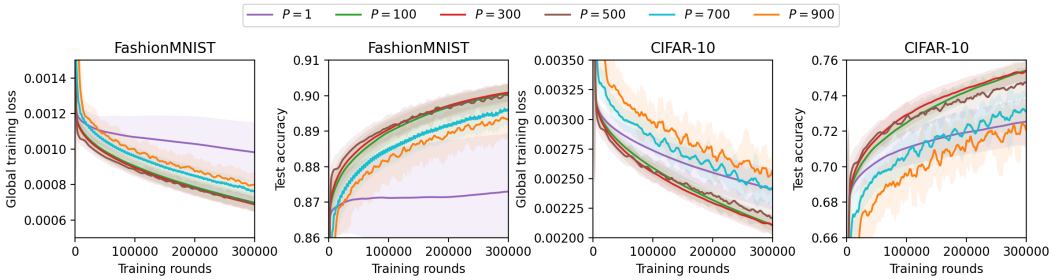


Figure 2: Algorithm 1 with amplification and different P, with periodically connected clients.

as described in the setup above. This corresponds to our setting with regularized participation. In this case, Algorithm 1 with amplification allows different groups of clients to make small progress among themselves first. After P rounds, each client has made its own share of contribution, and the updates get amplified so that the model parameter progresses faster towards the direction that is overall beneficial for all the clients. As our theory predicts, this approach gives a desirable convergence rate, which is now verified by experiments. Next, we observe that *the "wait-minibatch" and "wait-full" methods perform worse than Algorithm 1 with amplification*. We also observe that "wait-full", which computes the gradient on the entire dataset, does not provide substantial gain compared to "wait-minibatch" in the experiments. These observations align with our discussion in Section 5 on the practical performance of different methods. Finally, we observe that *standard FedAvg (i.e., Algorithm 1 without amplification) gives the lowest performance*. This is because without amplification, the algorithm cannot put more emphasis on the collective updates by the cohort of all clients, and the parameter updates may diverge from the overall optimal direction.

Different Values of P. Next, we fix $\eta = 10$ in Algorithm 1 and study the effect of choosing different values of P. We use the same learning rate γ that is obtained from grid search in the previous experiment. Note that the clients' availability pattern remains the same as described above. The choice of different P simulates the practical scenario where the estimation of P may not perfectly align with the actual participation cycle. The results are shown in Figure 2.

We observe that $P = 1$, which corresponds to the "classical" setting of FedAvg with two-sided learning rates [18, 36], does not give the best performance. Interestingly, compared to $P = 500$, we see that $P = 100$ and $P = 300$ give a similar performance, and even slightly better performance in the case of CIFAR-10 data. Due to the random offset applied to the first participation cycle in each experiment (see Appendix D.1 for details), every interval of 100 rounds can include the "partial" contributions by two subsets of clients (with data in 4, out of 10, majority classes). The results suggest that amplifying such partial contributions by multiple subsets of clients can still improve performance. In the case of CIFAR-10 data, the reason for $P = 500$ being slightly worse may be that the accumulated update within P rounds generally becomes larger as P gets larger, in which case amplification causes a bigger change in the model parameter. Depending on the landscape of the loss function, this change may be too big so that the model performance decreases. For a similar reason, the performances of $P = 700$ and $P = 900$ are even worse. Nevertheless, we expect that choosing a smaller learning rate γ can improve the performance for large P values.

7 Conclusion

In this paper, we have studied FL with arbitrary client participation. For a generalized FedAvg algorithm that amplifies parameter updates every P rounds, we have developed a unified framework for convergence analysis and obtained convergence rates for a variety of client participation patterns. Our findings suggest that regularized participation with finite P gives the best performance and matches the lower bound of SGD convergence error when T is sufficiently large. For stochastic participation, first, we have formally proven that convergence is guaranteed when the participation process is ergodic. Then, for two generic classes of participation processes, we have proven that with a properly chosen P, our convergence rate matches state-of-the-art FedAvg convergence rates that were derived for the idealized case of independent and unbiased participation. The empirical results have confirmed that amplification is useful and also provided further insights. Future directions include analysis of advanced FL algorithms and more detailed empirical study.

References

[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 2022.

[2] P. Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

[3] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019.

[4] W. Chen, S. Horvath, and P. Richtarik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

[5] Y. J. Cho, J. Wang, and G. Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, 2022.

[6] Y. Ding, C. Niu, Y. Yan, Z. Zheng, F. Wu, G. Chen, S. Tang, and R. Jia. Distributed optimization over block-cyclic data. *arXiv preprint arXiv:2002.07454*, 2020.

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

[8] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*, volume 139, pages 3407–3416. PMLR, Jul. 2021.

[9] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi. On the impact of client sampling on federated learning convergence. *arXiv preprint arXiv:2107.12211*, 2021.

[10] E. Gorbunov, F. Hanzely, and P. Richtarik. Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pages 3556–3564, 2021.

[11] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[12] X. Gu, K. Huang, J. Zhang, and L. Huang. Fast federated learning in the presence of arbitrary device unavailability. In *Advances in Neural Information Processing Systems*, 2021.

[13] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

[14] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[16] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[17] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[18] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[19] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

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

[22] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

[23] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations*, 2020.

[24] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[25] K. Mishchenko, A. Khaled Ragab Bayoumi, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

[26] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

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

[28] S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

[29] J. Wang and G. Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD. In *Proceedings of Machine Learning and Systems*, volume 1, pages 212–229, 2019.

[30] J. Wang and G. Joshi. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *Journal of Machine Learning Research*, 22(213):1–50, 2021.

[31] J. Wang, V. Tantia, N. Ballas, and M. Rabbat. Lookahead converges to stationary points of smooth non-convex functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8604–8608, 2020.

[32] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[33] B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

[34] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[35] Y. Yan, C. Niu, Y. Ding, Z. Zheng, F. Wu, G. Chen, S. Tang, and Z. Wu. Distributed non-convex optimization with sublinear speedup under intermittent client availability. *arXiv preprint arXiv:2002.07399*, 2020.

[36] H. Yang, M. Fang, and J. Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

[37] H. Yang, X. Zhang, P. Khanduri, and J. Liu. Anarchic federated learning. In *International Conference on Machine Learning*, pages 25331–25363. PMLR, 2022.

[38] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):12, 2019.

[39] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *International Conference on Machine Learning*, volume 97 of *PMLR*, pages 7184–7193, 2019.

[40] H. Yu, S. Yang, and S. Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *AAAI Conference on Artificial Intelligence*, pages 5693–5700, 2019.

[41] C. Yun, S. Rajput, and S. Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022.

[42] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems*, 32, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are described where necessary. More discussions are also available in the appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a foundational work that is widely applicable to various domains. The societal impact would depend on the specific application, which we believe is beyond the scope of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Included in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The pseudo-code of the algorithm is given in the main paper. The datasets and detailed experimental setup are described in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] More details are included in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The shaded areas in the figures represent the standard deviation. More details are explained in the appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Included in the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Mentioned in the appendix.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Additional Related Works **16**

B Additional Discussions **16**

B.1 Why Does Amplification Help? 16

B.2 Relating Different Participation Patterns to Practical FL Scenarios 17

C Proof **18**

C.1 Preliminaries 18

C.2 Lemmas 18

C.3 Proof of Theorem 3.1 24

C.4 Proof of Corollaries 3.2 and 3.3 29

D Additional Details and Results of Experiments **29**

D.1 Additional Details of the Setup 29

D.2 Additional Results 32

A Additional Related Works

We complement Section 1 by discussing a few additional related works as follows.

Client unavailability was also studied from the perspective of block-cyclic SGD [6, 7], which trains *multiple* models each for a block (group) of clients with known identity and was analyzed for convex objectives. Different from those works, our goal is to train a *single* model without requiring block structure, and we focus on non-convex objectives that frequently arise in modern problems involving neural networks.

In the literature of SGD in the centralized setting, the effect of different ways of minibatch sampling has also been studied, such as IID sampling from an arbitrary distribution [11] and random reshuffling [13, 25]. However, the problem in FL is much more complex and needs to incorporate the diversity of different clients' local objectives. Moreover, we focus on a unified framework that is more general than the special cases of IID sampling and random reshuffling.

FedAvg with two-sided learning rates has been studied in the literature [18, 36], where a global learning rate is applied to the updates at the end of each round, giving an improved convergence bound compared to using a single-sided learning rate. This is equivalent to our Algorithm 1 with $P = 1$ and the amplification factor η being the global learning rate. We extend this setup by allowing any $P \geq 1$. Moreover, most existing works including [18, 36] have only shown the benefit of two-sided learning rates in a theoretical setting; their experiments still uses a global learning rate of 1.0, which is equivalent to the case of a single-sided learning rate. Although some recent works have used different global learning rates in experiments, the best setting often remains to be choosing 1.0 as the global learning rate in the case of FedAvg [26, Table 8]. In contrast, we have verified the benefit of amplification in our experiments (Section 6), for scenarios with more challenging client participation patterns than existing works. The idea of amplification is also related to look-ahead algorithms [31, 42], which consider a two-agent setting, e.g., a server and a single client. Different from these works, we consider the unique challenges in federated learning with multiple arbitrarily participating clients, where the data is non-IID across clients.

It is also worth pointing out that our discussion related to waiting for all clients in Section 5 has some analogy to the comparison between local SGD and minibatch SGD [32, 33, 41]. Even when not considering the effect of partial participation, these existing works have only identified some specific classes of convex objectives where local SGD can be shown to have a better theoretical convergence rate than minibatch SGD. An improved theoretical approach to capture the fact that local SGD often outperforms minibatch SGD in practice, for a wide range of problems and objective functions (especially non-convex objectives), remains an interesting future direction.

B Additional Discussions

B.1 Why Does Amplification Help?

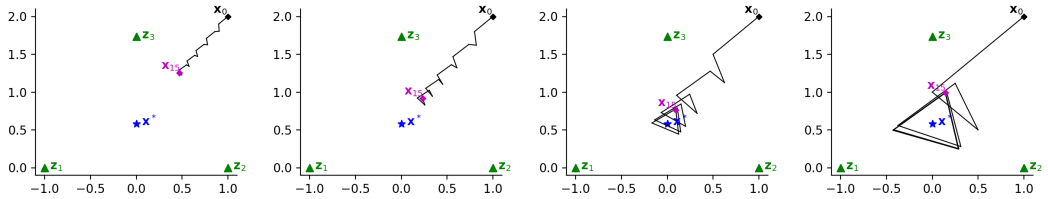
We give an illustrative example to explain why the use of amplification with $\eta > 1$ generally improves the performance, as shown in both our theory and experiments. Consider a simple setting with $N = 3$ clients. Each client n has a local objective defined as $F_n(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x} - \mathbf{z}_n\right\|^2$ for some constant vector $\mathbf{z}_n \in \mathbb{R}^m$. It is evident that the true optimal solution that minimizes the global objective $f(\mathbf{x}) = \frac{1}{6}\sum_{n=1}^{3}\left\|\mathbf{x} - \mathbf{z}_n\right\|^2$ is $\mathbf{x}^* = \frac{1}{3}\sum_{n=1}^{3}\mathbf{z}_n$, but we assume that this is not known to the system and we would like to solve this problem using Algorithm 1. In this specific example, we consider a two-dimensional space (i.e., $m = 2$) and define $\mathbf{z}_1 = (-1, 0), \mathbf{z}_2 = (1, 0), \mathbf{z}_3 = (0, \sqrt{3})$. We further assume that the three clients are available in a cyclic manner, so that in the first round, only client $n = 1$ is available; in the second round, only client $n = 2$ is available; in the third round, only client $n = 3$ is available; then this cycle continues with client $n = 1$ available in the fourth round, and so on. We plot the trajectory of how the model parameter $\mathbf{x}$ changes from the initial value $\mathbf{x}_0$ to $\mathbf{x}_{15}$ that is obtained after $T = 15$ rounds. The initial parameter is set to $\mathbf{x}_0 = (1, 2)$ in this example. The plots of the trajectories with different choices of local learning rate γ and amplification factor η are shown in Figures B.1–B.2.

With the local objectives defined in this example, the local gradient can be directly computed as $\nabla F_n(\mathbf{x}) = \mathbf{x} - \mathbf{z}_n$. Therefore, in the case of no amplification (i.e., $\eta = 1$), the solution variable $\mathbf{x}$ moves towards $\mathbf{z}_1$ in the first round when client $n = 1$ is available; then, it moves towards $\mathbf{z}_2$ in the

second round when client $n = 2$ is available, and so on. As we see from Figure B.1, this can lead to a pattern where $\mathbf{x}$ cycles around the optimal solution $\mathbf{x}^*$ but approaches $\mathbf{x}^*$ only very slowly, especially when γ is large. In the case of small γ, this cyclic pattern is not apparent, but we can still see that $\mathbf{x}$ moves towards different local optimal points (i.e., $\mathbf{z}_n$ for different n) in different rounds, and the overall convergence speed is slow. This leads to the following problem: regardless of whether we choose a large or small γ, at the end of the 15-th round, there is an apparent gap between $\mathbf{x}_{15}$ and $\mathbf{x}^*$.

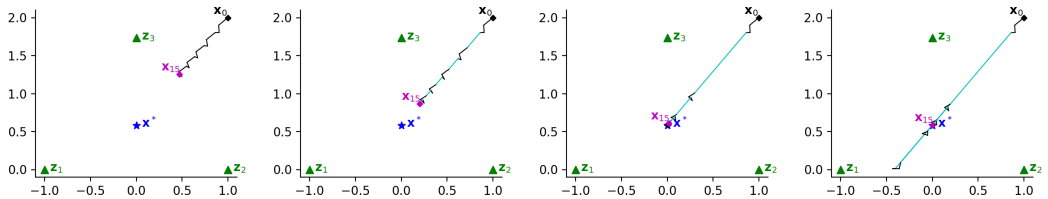
The advantage of amplification is that it increases the importance of the aggregated updates made by clients that participate in different rounds. In this specific example, a single client's update direction may be towards its own local optimum, which can be different from the direction towards the global optimum. However, the accumulated updates by all the three clients over every three rounds will be more likely towards the global optimum (or a neighborhood of it). By amplifying such accumulated updates after every three rounds (i.e., $P = 3$), we can let the solution variable $\mathbf{x}$ move much faster towards the true (global) optimum. This intuition is confirmed by the trajectories shown in Figure B.2. Here, we note that the product of γ and η in the corresponding sub-figures of Figures B.1 and B.2 are equal. By keeping a small local learning rate γ and amplifying the updates accumulated in every $P = 3$ rounds, we can converge to the optimal value $\mathbf{x}^*$, as shown in Figure B.2.

Note that in practice, P does not need to perfectly align with the "cycle" of participation, as seen in Figure 2 in Section 6 of the main paper.



(a) $\gamma = 0.05, \eta = 1$ (b) $\gamma = 0.1, \eta = 1$ (c) $\gamma = 0.25, \eta = 1$ (d) $\gamma = 0.5, \eta = 1$

Figure B.1: Motivating example: different local learning rates γ without amplification (i.e., $\eta = 1$). The trajectory from $\mathbf{x}_0$ to $\mathbf{x}_{15}$ shows how the model parameter changes from round $t = 0$ to round $t = 15$.



(a) $\gamma = 0.05, \eta = 1$ (b) $\gamma = 0.05, \eta = 2$ (c) $\gamma = 0.05, \eta = 5$ (d) $\gamma = 0.05, \eta = 10$

Figure B.2: Motivating example: fixed local learning rate $\gamma = 0.05$ with different amplification factors η. The trajectory from $\mathbf{x}_0$ to $\mathbf{x}_{15}$ shows how the model parameter changes from round $t = 0$ to round $t = 15$. The segments in cyan color shows the change in parameter $\mathbf{x}$ due to amplification, while the segments in black color shows the parameter change due to regular SGD operation.

B.2 Relating Different Participation Patterns to Practical FL Scenarios

In the following, we discuss some possible connections between practical FL application scenarios and the different participation patterns that are considered in our theoretical analysis.

The case of periodic shifts, which is often seen in practical cross-device FL scenarios, is mostly related to *regularized participation*. In this case, the clients participate equally over the entire period, but only a subset of clients from a specific population (and with a specific data distribution) may participate in each "period" that includes a certain number of consecutive rounds. In practice, this "equal participation" only needs to hold approximately, and our empirical results at the end of

Section 6 show that our Algorithm 1 can still give good performance even if P is less than the full cycle (period) of participation.

Among the stochastic participation patterns, *ergodic* is the most generic. Intuitively, it says that the participation weights of different clients are the same when averaged over a long enough time. This can represent a cross-device FL scenario, where the participation of each client at any time instance can be highly random, but the long-term average statistics of clients are the same. The case of *strongly mixing* participation includes the special case where the participation follows a Markov chain, i.e., if the client is currently unavailable, there is a certain (possibly small) probability that it will become available in the next round, and vice versa. This can also represent a cross-device FL setting, where the devices get connected and disconnected over time, and there is some randomness in the exact time it gets connected or disconnected. Finally, the case of *independent* participation may represent a cross-silo FL setting, where the reason for not participating in a certain round is not because the client is disconnected for an extended period of time, but rather because the client has some other higher-priority tasks to run so that it cannot participate in all rounds.

C Proof

C.1 Preliminaries

The following inequalities are frequently used throughout our proofs. We will use them without further explanation.

From Jensen's inequality, for any $\mathbf{z}_m \in \mathbb{R}^d, m \in \{1, 2, \ldots, M\}$, we have

$$\left\| \frac{1}{M} \sum_{m=1}^{M} \mathbf{z}_m \right\|^2 \leq \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{z}_m\|^2 \tag{C.1}$$

which directly gives

$$\left\| \sum_{m=1}^{M} \mathbf{z}_m \right\|^2 \leq M \sum_{m=1}^{M} \|\mathbf{z}_m\|^2 . \tag{C.2}$$

Peter-Paul inequality (also known as Young's inequality) gives

$$\langle \mathbf{z}_1, \mathbf{z}_2 \rangle \leq \frac{b \|\mathbf{z}_1\|^2}{2} + \frac{\|\mathbf{z}_2\|^2}{2b} \tag{C.3}$$

for any $b > 0$ and any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$. Hence, we also have

$$\|\mathbf{z}_1 + \mathbf{z}_2\|^2 \leq (1 + b) \|\mathbf{z}_1\|^2 + \left(1 + \frac{1}{b}\right) \|\mathbf{z}_2\|^2 . \tag{C.4}$$

Throughout the analysis, we use the short-hand notion $\mathbb{E}_{\cdot|\mathcal{Q}}[\cdot]$ to denote $\mathbb{E}[\cdot|\mathcal{Q}]$.

In addition, we use the notations in Algorithm 1.

C.2 Lemmas

Lemma C.1. *When $\gamma \leq \frac{1}{12LIP}$, we have*

$$\sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \leq 3I\gamma^2 \left(6I\tilde{\nu}^2 + \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \right) . \tag{C.5}$$

Proof. Note that

$$\sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i+1}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i+1}^{n'} \right\|^2 \right]$$

$$= \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \gamma \mathbf{g}_n(\mathbf{y}_{t,i}^n) - \sum_{n'=1}^{N} q_t^{n'} (\mathbf{y}_{t,i}^{n'} - \gamma \mathbf{g}_{n'}(\mathbf{y}_{t,i}^{n'})) \right\|^2 \right]$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \gamma \nabla F_n(\mathbf{y}_{t,i}^n) - \sum_{n'=1}^{N} q_t^{n'} (\mathbf{y}_{t,i}^{n'} - \gamma \nabla F_{n'}(\mathbf{y}_{t,i}^{n'})) \right\|^2 \right] \\
&\quad + \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \gamma \left(\mathbf{g}_n(\mathbf{y}_{t,i}^n) - \nabla F_n(\mathbf{y}_{t,i}^n) \right) - \gamma \sum_{n'=1}^{N} q_t^{n'} \left(\mathbf{g}_{n'}(\mathbf{y}_{t,i}^{n'}) - \nabla F_{n'}(\mathbf{y}_{t,i}^{n'}) \right) \right\|^2 \right] \\
&= \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \gamma \nabla F_n(\mathbf{y}_{t,i}^n) - \sum_{n'=1}^{N} q_t^{n'} (\mathbf{y}_{t,i}^{n'} - \gamma \nabla F_{n'}(\mathbf{y}_{t,i}^{n'})) \right\|^2 \right] \\
&\quad + \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \Bigg[\Bigg\| - \gamma \sum_{n' \in \{1,\dots,N\} \setminus \{n\}} q_t^{n'} \left(\mathbf{g}_{n'}(\mathbf{y}_{t,i}^{n'}) - \nabla F_{n'}(\mathbf{y}_{t,i}^{n'}) \right) \\
&\qquad\qquad - \gamma \left(q_t^n - 1 \right) \left(\mathbf{g}_n(\mathbf{y}_{t,i}^n) - \nabla F_n(\mathbf{y}_{t,i}^n) \right) \Bigg\|^2 \Bigg] \\
&\overset{(b)}{\leq} \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \gamma \nabla F_n(\mathbf{y}_{t,i}^n) - \sum_{n'=1}^{N} q_t^{n'} (\mathbf{y}_{t,i}^{n'} - \gamma \nabla F_{n'}(\mathbf{y}_{t,i}^{n'})) \right\|^2 \right] + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
&= \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \Bigg[\Bigg\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} - \gamma \Bigg[\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \\
&\qquad\qquad + \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \\
&\qquad\qquad + \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'}(\mathbf{y}_{t,i}^{n'}) \Bigg] \Bigg\|^2 \Bigg] \\
&\quad + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
&\leq \left(1 + \frac{1}{2I-1} \right) \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \\
&\quad + 2I\gamma^2 \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \Bigg[\Bigg\| \nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) + \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \\
&\qquad\qquad - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) + \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \\
&\qquad\qquad - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'}(\mathbf{y}_{t,i}^{n'}) \Bigg\|^2 \Bigg] \\
&\quad + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
&\leq \left(1 + \frac{1}{2I-1} \right) \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \\
&\quad + 6I\gamma^2 \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \right\|^2 \right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 6I\gamma^2 \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \nabla F_n \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) \right\|^2 \right] \\
&+ 6I\gamma^2 \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'} \left(\sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right) - \sum_{n'=1}^{N} q_t^{n'} \nabla F_{n'}(\mathbf{y}_{t,i}^{n'}) \right\|^2 \right] \\
&+ \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
\overset{(c)}{\leq} &\left(1 + \frac{1}{2I-1} \right) \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \\
&+ 6IL^2\gamma^2 \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right\|^2 \right] + 6I\gamma^2 \tilde{\nu}^2 \\
&+ 6IL^2\gamma^2 \sum_{n=1}^{N} q_t^n \sum_{n'=1}^{N} q_t^{n'} \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^{n'} - \sum_{n''=1}^{N} q_t^{n''} \mathbf{y}_{t,i}^{n''} \right\|^2 \right] + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
\overset{(d)}{=} &\left(1 + \frac{1}{2I-1} + 12IL^2\gamma^2 \right) \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \\
&+ 6I\gamma^2 \tilde{\nu}^2 + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right) \\
\overset{(e)}{\leq} &\left(1 + \frac{7}{12\left(I - \frac{1}{2}\right)} \right) \sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] + 6I\gamma^2 \tilde{\nu}^2 + \gamma^2 \sigma^2 \left(1 - \sum_{n=1}^{N} (q_t^n)^2 \right).
\end{aligned}
\tag{C.6}
$$

In the above, (a) is due to $\mathbb{E}_{\cdot|\mathcal{Q}} \left[\|\mathbf{z}\|^2 \right] = \left\| \mathbb{E}_{\cdot|\mathcal{Q}} [\mathbf{z}] \right\|^2 + \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{z} - \mathbb{E}_{\cdot|\mathcal{Q}} [\mathbf{z}] \right\|^2 \right]$; (b) is because the stochastic gradient noise is independent across clients, the variance of the sum of independent random variables is equal to the sum of the variance and $\mathrm{Var}(a\mathbf{z}) = a^2 \mathrm{Var}(\mathbf{z})$ where $\mathrm{Var}(\mathbf{z}) := \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{z} - \mathbb{E}_{\cdot|\mathcal{Q}} [\mathbf{z}] \right\|^2 \right]$, and also

$$
\begin{aligned}
&\sum_{n=1}^{N} q_t^n \left[\sum_{n' \in \{1,\ldots,N\} \setminus \{n\}} (q_t^{n'})^2 + (1 - q_t^n)^2 \right] = \sum_{n=1}^{N} q_t^n \left[\sum_{n'=1}^{N} (q_t^{n'})^2 - (q_t^n)^2 + (1 - q_t^n)^2 \right] \\
&= \sum_{n=1}^{N} (q_t^n)^2 + \sum_{n=1}^{N} q_t^n (1 - 2q_t^n) = \sum_{n=1}^{N} (q_t^n)^2 + \sum_{n=1}^{N} q_t^n - 2 \sum_{n=1}^{N} (q_t^n)^2 = 1 - \sum_{n=1}^{N} (q_t^n)^2;
\end{aligned}
$$

(c) uses smoothness in the second term, the definition of $\tilde{\nu}$ in the third term, and Jensen's inequality followed by smoothness in the fourth term, while noting that $\sum_{n=1}^{N} q_t^n = 1$; (d) combines the second and fourth terms in the above line, while (again) noting that $\sum_{n=1}^{N} q_t^n = 1$; (e) is because

$$
\frac{1}{2I-1} + 12IL^2\gamma^2 \leq \frac{1}{2I-1} + \frac{1}{12I} \leq \frac{1}{2\left(I - \frac{1}{2}\right)} + \frac{1}{12\left(I - \frac{1}{2}\right)} = \frac{7}{12\left(I - \frac{1}{2}\right)}
$$

where the first inequality is due to $\gamma \leq \frac{1}{12LIP} \leq \frac{1}{12LI}$ since $P \geq 1$.

For $i = 0$, we note that

$$
\sum_{n=1}^{N} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,0}^n - \sum_{n'=1}^{N} q_t^{n'} \mathbf{y}_{t,0}^{n'} \right\|^2 \right] = 0 \tag{C.7}
$$

because $\mathbf{y}_{t,0}^n = \mathbf{x}_t$ for all n as specified in Algorithm 1, and $\sum_{n=1}^N q_t^n = 1$.
By combining (C.6) and (C.7) and unrolling the recursion, we have for any i that

$$
\begin{aligned}
&\sum_{n=1}^N q_t^n \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \mathbf{y}_{t,i}^n - \sum_{n'=1}^N q_t^{n'} \mathbf{y}_{t,i}^{n'} \right\|^2 \right] \\
&\leq \sum_{i=0}^{I-1} \left(1 + \frac{7}{12\left(I - \frac{1}{2}\right)} \right)^i \left(6I\gamma^2\tilde{\nu}^2 + \gamma^2\sigma^2 \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right) \\
&\overset{(a)}{=} \left[\left(1 + \frac{7}{12\left(I - \frac{1}{2}\right)} \right)^I - 1 \right] \cdot \frac{12\left(I - \frac{1}{2}\right)}{7} \cdot \left(6I\gamma^2\tilde{\nu}^2 + \gamma^2\sigma^2 \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right) \\
&= \left[\left(1 + \frac{7}{12\left(I - \frac{1}{2}\right)} \right)^{I - \frac{1}{2}} \left(1 + \frac{7}{12\left(I - \frac{1}{2}\right)} \right)^{\frac{1}{2}} - 1 \right] \cdot \frac{12\left(I - \frac{1}{2}\right)}{7} \cdot \left(6I\gamma^2\tilde{\nu}^2 + \gamma^2\sigma^2 \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right) \\
&\overset{(b)}{\leq} \left[e^{\frac{7}{12}} \cdot \sqrt{\frac{13}{6}} - 1 \right] \cdot \frac{12\left(I - \frac{1}{2}\right)}{7} \cdot \left(6I\gamma^2\tilde{\nu}^2 + \gamma^2\sigma^2 \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right) \\
&\overset{(c)}{\leq} 3I\gamma^2 \left(6I\tilde{\nu}^2 + \sigma^2 \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right)
\end{aligned}
$$

where (a) uses the expression of the sum of geometric progression series; (b) uses $(1+x)^{\frac{1}{x}} \leq e$ and $I \geq 1$ hence $\left(1 + \frac{7}{12\left(I-\frac{1}{2}\right)}\right)^{\frac{1}{2}} \leq \left(1 + \frac{7}{6}\right)^{\frac{1}{2}} = \sqrt{\frac{13}{6}}$; (c) is because $\left[e^{\frac{7}{12}} \cdot \sqrt{\frac{13}{6}} - 1\right] \frac{12}{7} \leq 3$ and $I - \frac{1}{2} \leq I$. □

Lemma C.2. *When $\gamma \leq \frac{1}{12LIP}$, we have*

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0} \right\|^2 \right] \\
&\leq 3\gamma^2 I^2 \tilde{\nu}^2 + 24\gamma^2 I^2 P^2 \tilde{\beta}^2 + 24\gamma^2 I^2 P^2 \mathbb{E}_{\cdot|\mathcal{Q}} \left[\|\nabla f(\mathbf{x}_{t_0})\|^2 \right] \\
&\quad + 3\gamma^2 IP\sigma^2 \left[\rho^2 + \frac{1}{6P} \left(1 - \sum_{n=1}^N (q_t^n)^2 \right) \right]
\end{aligned} \tag{C.8}
$$

for $t_0 \leq t \leq t_0 + P - 1$.

Proof. We have

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \mathbf{y}_{t,i+1}^n - \mathbf{x}_{t_0} \right\|^2 \right] \\
&= \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \left(\mathbf{y}_{t,i}^n - \gamma \mathbf{g}_n(\mathbf{y}_{t,i}^n) \right) - \mathbf{x}_{t_0} \right\|^2 \right] \\
&= \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \left(\mathbf{y}_{t,i}^n - \gamma \nabla F_n(\mathbf{y}_{t,i}^n) \right) - \mathbf{x}_{t_0} \right\|^2 \right] + \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \gamma \left(\mathbf{g}_n(\mathbf{y}_{t,i}^n) - \nabla F_n(\mathbf{y}_{t,i}^n) \right) \right\|^2 \right] \\
&\leq \mathbb{E}_{\cdot|\mathcal{Q}} \left[\left\| \sum_{n=1}^N q_t^n \left[\mathbf{y}_{t,i}^n - \gamma \left(\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n \left(\sum_{n'=1}^N q_t^{n'} \mathbf{y}_{t,i}^{n'} \right) + \nabla F_n \left(\sum_{n'=1}^N q_t^{n'} \mathbf{y}_{t,i}^{n'} \right) \right. \right. \right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \nabla F_n(\mathbf{x}_{t_0}) + \nabla F_n(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0}) + \nabla f(\mathbf{x}_{t_0})\Bigg)\Bigg] - \mathbf{x}_{t_0}\Bigg\|^2\Bigg] + \gamma^2\rho^2\sigma^2 \\
&\leq \left(1 + \frac{1}{2IP-1}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n\left[\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n\left(\sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'}\right)\right]\right\|^2\right] \\
&\quad + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n\left[\nabla F_n\left(\sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'}\right) - \nabla F_n(\mathbf{x}_{t_0})\right]\right\|^2\right] \\
&\quad + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n\left[\nabla F_n(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0})\right]\right\|^2\right] \\
&\quad + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n\nabla f(\mathbf{x}_{t_0})\right\|^2\right] + \gamma^2\rho^2\sigma^2 \\
&\overset{(a)}{\leq} \left(1 + \frac{1}{2IP-1}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + 8IP\gamma^2\sum_{n=1}^N q_t^n\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n\left(\sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'}\right)\right\|^2\right] \\
&\quad + 8IP\gamma^2\sum_{n=1}^N q_t^n\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\nabla F_n\left(\sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'}\right) - \nabla F_n(\mathbf{x}_{t_0})\right\|^2\right] \\
&\quad + 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\rho^2\sigma^2 \\
&\overset{(b)}{\leq} \left(1 + \frac{1}{2IP-1}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + 8IPL^2\gamma^2\sum_{n=1}^N q_t^n\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{y}_{t,i}^n - \sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'}\right\|^2\right] \\
&\quad + 8IPL^2\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n'=1}^N q_t^{n'}\mathbf{y}_{t,i}^{n'} - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\rho^2\sigma^2 \\
&\overset{(c)}{\leq} \left(1 + \frac{1}{2IP-1} + 8IPL^2\gamma^2\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + 24I^2PL^2\gamma^4\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^N (q_t^n)^2\right)\right) \\
&\quad + 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\rho^2\sigma^2 \\
&\overset{(d)}{\leq} \left(1 + \frac{5}{9\left(IP - \frac{1}{2}\right)}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^N q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] + 144\gamma^4 I^3PL^2\tilde{\nu}^2
\end{aligned}
$$

$$+ 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\sigma^2\left[\rho^2 + 24\gamma^2 I^2 PL^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]$$

$$\overset{(e)}{\leq} \left(1 + \frac{5}{9\left(IP - \frac{1}{2}\right)}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right] + \frac{\gamma^2 I\tilde{\nu}^2}{P}$$

$$+ 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]. \tag{C.9}$$

In the above, the stochastic gradient variance decomposition is the same as the proof of Lemma C.1, while noting that $\sum_{n=1}^{N}(q_t^n)^2 \leq \rho^2$; (a) uses Jensen's inequality, the definition of $\tilde{\beta}^2$, and $\sum_{n=1}^{N} q_t^n = 1$; (b) uses L-smoothness; (c) is from Lemma C.1; (d) is due to

$$\frac{1}{2IP-1} + 8IPL^2\gamma^2 \leq \frac{1}{2IP-1} + \frac{1}{18IP} \leq \frac{1}{2\left(IP - \frac{1}{2}\right)} + \frac{1}{18\left(IP - \frac{1}{2}\right)} = \frac{5}{9\left(IP - \frac{1}{2}\right)}$$

where the first inequality is due to $\gamma \leq \frac{1}{12LIP}$; (e) uses $\gamma \leq \frac{1}{12LIP}$ in the second and last terms.

We note that, for $t_0 \leq t \leq t_0 + P - 2$, we have

$$\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_{t+1}^n \mathbf{y}_{t+1,0}^n - \mathbf{x}_{t_0}\right\|^2\right]$$
$$= \mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_{t+1}^n \mathbf{x}_{t+1} - \mathbf{x}_{t_0}\right\|^2\right] = \mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_{t_0}\|^2\right] = \mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_t^n \mathbf{y}_{t,I}^n - \mathbf{x}_{t_0}\right\|^2\right] \tag{C.10}$$

because $\mathbf{y}_{t,0}^n = \mathbf{x}_t$ for all n as specified in Algorithm 1, and $\sum_{n=1}^{N} q_t^n = 1$ hence $\mathbf{x}_{t+1} = \mathbf{x}_t + \sum_{n=1}^{N} q_t^n(\mathbf{y}_{t,I}^n - \mathbf{x}_t) = \sum_{n=1}^{N} q_t^n \mathbf{y}_{t,I}^n$ according to Algorithm 1.

Therefore, the same recursion in (C.9) also holds between $t+1$ and t, i.e.,

$$\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_{t+1}^n \mathbf{y}_{t+1,1}^n - \mathbf{x}_{t_0}\right\|^2\right]$$
$$\leq \left(1 + \frac{5}{9\left(IP - \frac{1}{2}\right)}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_t^n \mathbf{y}_{t,I}^n - \mathbf{x}_{t_0}\right\|^2\right]$$
$$+ \frac{\gamma^2 I\tilde{\nu}^2}{P} + 8IP\gamma^2\tilde{\beta}^2 + 8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] + \gamma^2\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]. \tag{C.11}$$

When $t = t_0$ and $i = 0$, note that

$$\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_{t_0}^n \mathbf{y}_{t_0,0}^n - \mathbf{x}_{t_0}\right\|^2\right] = 0 \tag{C.12}$$

because $\mathbf{y}_{t_0,0}^n = \mathbf{x}_{t_0}$ for all n as specified in Algorithm 1, and $\sum_{n=1}^{N} q_t^n = 1$.

We combine (C.9), (C.11), and (C.12). Since $t_0 \leq t \leq t_0 + P - 1$, unrolling the recursion gives

$$\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n=1}^{N} q_t^n \mathbf{y}_{t,i}^n - \mathbf{x}_{t_0}\right\|^2\right]$$

$$
\begin{aligned}
&\leq \sum_{\kappa=0}^{IP-1}\left(1+\frac{5}{9\left(IP-\frac{1}{2}\right)}\right)^{\kappa}\left(\frac{\gamma^2 I\tilde{\nu}^2}{P}+8IP\gamma^2\tilde{\beta}^2+8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]\right.\\
&\qquad\qquad\left.+\gamma^2\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)\\
&=\left[\left(1+\frac{5}{9\left(IP-\frac{1}{2}\right)}\right)^{IP}-1\right]\cdot\frac{9\left(IP-\frac{1}{2}\right)}{5}\cdot\left(\frac{\gamma^2 I\tilde{\nu}^2}{P}+8IP\gamma^2\tilde{\beta}^2+8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]\right.\\
&\qquad\qquad\left.+\gamma^2\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)\\
&=\left[\left(1+\frac{5}{9\left(IP-\frac{1}{2}\right)}\right)^{IP-\frac{1}{2}}\left(1+\frac{5}{9\left(IP-\frac{1}{2}\right)}\right)^{\frac{1}{2}}-1\right]\cdot\frac{9\left(IP-\frac{1}{2}\right)}{5}\cdot\left(\frac{\gamma^2 I\tilde{\nu}^2}{P}+8IP\gamma^2\tilde{\beta}^2\right.\\
&\qquad\left.+8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]+\gamma^2\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)\\
&\leq\left[e^{\frac{5}{9}}\cdot\sqrt{\frac{19}{9}}-1\right]\cdot\frac{9\left(IP-\frac{1}{2}\right)}{5}\cdot\left(\frac{\gamma^2 I\tilde{\nu}^2}{P}+8IP\gamma^2\tilde{\beta}^2+8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]\right.\\
&\qquad\qquad\left.+\gamma^2\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)\\
&\leq 3IP\left(\frac{\gamma^2 I\tilde{\nu}^2}{P}+8IP\gamma^2\tilde{\beta}^2+8IP\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]+\gamma^2\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)\\
&\leq 3\gamma^2 I^2\tilde{\nu}^2+24\gamma^2 I^2P^2\tilde{\beta}^2+24\gamma^2 I^2P^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]+3\gamma^2 IP\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]
\end{aligned}
$$

where the rationale behind the steps is similar to the proof of Lemma C.1. □

C.3 Proof of Theorem 3.1

Consider $t_0 = kP$ for $k = 0, 1, 2, ...$

Due to smoothness,

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[f(\mathbf{x}_{t_0+P})\right]\\
&\leq f(\mathbf{x}_{t_0})+\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\langle\nabla f(\mathbf{x}_{t_0}),\mathbf{x}_{t_0+P}-\mathbf{x}_{t_0}\rangle\right]+\frac{L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\|\mathbf{x}_{t_0+P}-\mathbf{x}_{t_0}\|^2\right]\\
&\leq f(\mathbf{x}_{t_0})-\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}),\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right]\right\rangle\\
&\qquad+\frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\overset{(a)}{=} f(\mathbf{x}_{t_0})-\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}),\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbb{E}\left[\mathbf{g}_n(\mathbf{y}_{t,i}^n)\middle|\mathcal{Q},\mathbf{y}_{t,i}^n,\mathbf{x}_{t_0}\right]\right]\right\rangle\\
&\qquad+\frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\overset{(b)}{=} f(\mathbf{x}_{t_0})-\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}),\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right]\right\rangle
\end{aligned}
$$

$$+ \frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]$$

where $\mathbb{E}_{\cdot|\mathcal{Q},t_0}[\mathbf{z}]$ is a short-hand notion for $\mathbb{E}\left[\mathbf{z}\middle|\mathcal{Q},\mathbf{x}_{t_0}\right]$; (a) is due to the law of total expectation, where the expectation is taken over $\mathbf{y}_{t,i}^n$; (b) is due to the unbiasedness and independence of stochastic gradients. In other parts of the proof, we may use the law of total expectation in a similar way without explanation.

Taking expectation on both sides over $\mathbf{x}_{t_0}$, we obtain

$$\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right] \leq \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}), \sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\rangle\right]$$
$$+ \frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]. \tag{C.13}$$

Consider the second term in (C.13),

$$\begin{aligned}
&-\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}), \sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\rangle \\
&= -\frac{\gamma\eta}{IP}\left\langle IP\nabla f(\mathbf{x}_{t_0}), \sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\rangle \\
&\overset{(a)}{=} \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n) - IP\nabla f(\mathbf{x}_{t_0})\right\|^2 \\
&\quad - \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2 \\
&= \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\left[\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla f(\mathbf{x}_{t_0})\right]\right\|^2 \\
&\quad - \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2 \\
&= \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\left[\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right) + \nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right)\right.\right. \\
&\qquad\qquad \left.\left. - \nabla F_n(\mathbf{x}_{t_0}) + \nabla F_n(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0})\right]\right\|^2 \\
&\quad - \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2 \\
&\overset{(b)}{\leq} \frac{3\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\left[\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right)\right]\right\|^2 \\
&\quad + \frac{3\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\left[\nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right) - \nabla F_n(\mathbf{x}_{t_0})\right]\right\|^2 \\
&\quad + \frac{3\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n\sum_{i=0}^{I-1}\left[\nabla F_n(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0})\right]\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&- \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2 \\
&\overset{(c)}{\leq} \frac{3\gamma\eta}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}\sum_{i=0}^{I-1} q_t^n \left\|\nabla F_n(\mathbf{y}_{t,i}^n) - \nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right)\right\|^2 \\
&\quad + \frac{3\gamma\eta}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}\sum_{i=0}^{I-1} q_t^n \left\|\nabla F_n\left(\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right) - \nabla F_n(\mathbf{x}_{t_0})\right\|^2 \\
&\quad + \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2 \\
&\leq \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}\sum_{i=0}^{I-1} q_t^n \left\|\mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right\|^2 \\
&\quad + \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{i=0}^{I-1}\left\|\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'} - \mathbf{x}_{t_0}\right\|^2 \\
&\quad + \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2 - \frac{\gamma\eta}{2IP}\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2
\end{aligned}
$$

where (a) uses $-\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}(\|\mathbf{a}-\mathbf{b}\|^2 - \|\mathbf{a}\|^2 - \|\mathbf{b}\|^2)$; (b) uses Jensen's inequality in the first term; (c) uses Jensen's inequality in the first two terms and the definition of $\tilde{\delta}^2(P)$ in the third term. Hence,

$$
\begin{aligned}
&- \mathbb{E}_{\cdot|\mathcal{Q}}\left[\gamma\eta\left\langle \nabla f(\mathbf{x}_{t_0}), \sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\rangle\right] \\
&\leq \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}\sum_{i=0}^{I-1} q_t^n \mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{y}_{t,i}^n - \sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'}\right\|^2\right] \\
&\quad + \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{i=0}^{I-1}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{n'=1}^{N} q_t^{n'}\mathbf{y}_{t,i}^{n'} - \mathbf{x}_{t_0}\right\|^2\right] \\
&\quad + \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2\right] - \frac{\gamma\eta}{2IP}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] \\
&\overset{(a)}{\leq} \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{i=0}^{I-1} 3I\gamma^2\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right) \\
&\quad + \frac{3\gamma\eta L^2}{2}\sum_{t=t_0}^{t_0+P-1}\sum_{i=0}^{I-1}\Bigg(3\gamma^2 I^2\tilde{\nu}^2 + 24\gamma^2 I^2 P^2\tilde{\beta}^2 + 24\gamma^2 I^2 P^2 \mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2\right] \\
&\qquad\qquad + 3\gamma^2 IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]\Bigg) \\
&\quad + \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\nabla f(\mathbf{x}_{t_0})\right\|^2\right] - \frac{\gamma\eta}{2IP}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] \\
&= \frac{9\gamma^3\eta L^2 I^2 P}{2}\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right)
\end{aligned}
$$

$$+\frac{3\gamma\eta L^2 IP}{2}\Bigg(3\gamma^2 I^2\tilde{\nu}^2+24\gamma^2 I^2P^2\tilde{\beta}^2+24\gamma^2 I^2P^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]$$
$$+3\gamma^2 IP\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\Bigg)$$
$$+\frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2}-\frac{\gamma\eta IP}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]-\frac{\gamma\eta}{2IP}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] \tag{C.14}$$

where (a) uses Lemmas C.1 and C.2.

Note also that

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\overset{(a)}{=}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\quad+\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\left[\mathbf{g}_n(\mathbf{y}_{t,i}^n)-\nabla F_n(\mathbf{y}_{t,i}^n)\right]\right\|^2\right]\\
&\overset{(b)}{=}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\quad+\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}(q_t^n)^2\sum_{i=0}^{I-1}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{g}_n(\mathbf{y}_{t,i}^n)-\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\overset{(c)}{\leq}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right]+IP\rho^2\sigma^2
\end{aligned} \tag{C.15}$$

where (a) uses $\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\mathbf{z}\|^2\right]=\left\|\mathbb{E}_{\cdot|\mathcal{Q}}[\mathbf{z}]\right\|^2+\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{z}-\mathbb{E}_{\cdot|\mathcal{Q}}[\mathbf{z}]\right\|^2\right]$; (b) is because the variance of the sum of independent random variables is equal to the sum of the variance and $\mathrm{Var}(a\mathbf{z})=a^2\mathrm{Var}(\mathbf{z})$ where $\mathrm{Var}(\mathbf{z}):=\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{z}-\mathbb{E}_{\cdot|\mathcal{Q}}[\mathbf{z}]\right\|^2\right]$; (c) uses the definitions of σ^2 and ρ^2.

Plugging (C.14) and (C.15) back into (C.13), we get

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right]\\
&\leq\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right]-\mathbb{E}_{\cdot|\mathcal{Q}}\left[\gamma\eta\left\langle\nabla f(\mathbf{x}_{t_0}),\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\rangle\right]\\
&\quad+\frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N}q_t^n\sum_{i=0}^{I-1}\mathbf{g}_n(\mathbf{y}_{t,i}^n)\right\|^2\right]\\
&\leq\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right]+\frac{9\gamma^3\eta L^2I^2P}{2}\left(6I\tilde{\nu}^2+\sigma^2\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right)\\
&\quad+\frac{3\gamma\eta L^2IP}{2}\Bigg(3\gamma^2I^2\tilde{\nu}^2+24\gamma^2I^2P^2\tilde{\beta}^2+24\gamma^2I^2P^2\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right]\\
&\qquad+3\gamma^2IP\sigma^2\left[\rho^2+\frac{1}{6P}\left(1-\sum_{n=1}^{N}(q_t^n)^2\right)\right]\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] - \frac{\gamma\eta}{2IP}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] \\
&+ \frac{\gamma^2\eta^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] + \frac{\gamma^2\eta^2 LIP\rho^2\sigma^2}{2} \\
=& \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] + \frac{9\gamma^3\eta L^2 I^2 P}{2}\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right) \\
&+ \frac{3\gamma\eta L^2 IP}{2}\left(3\gamma^2 I^2\tilde{\nu}^2 + 24\gamma^2 I^2 P^2\tilde{\beta}^2 + 3\gamma^2 IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]\right) \\
&+ \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{2}\left(1 - 72\gamma^2 L^2 I^2 P^2\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&- \left(\frac{\gamma\eta}{2IP} - \frac{\gamma^2\eta^2 L}{2}\right)\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\sum_{t=t_0}^{t_0+P-1}\sum_{n=1}^{N} q_t^n \sum_{i=0}^{I-1}\nabla F_n(\mathbf{y}_{t,i}^n)\right\|^2\right] \\
&+ \frac{\gamma^2\eta^2 LIP\rho^2\sigma^2}{2} \\
\leq& \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] + \frac{9\gamma^3\eta L^2 I^2 P}{2}\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right) \\
&+ \frac{3\gamma\eta L^2 IP}{2}\left(3\gamma^2 I^2\tilde{\nu}^2 + 24\gamma^2 I^2 P^2\tilde{\beta}^2 + 3\gamma^2 IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]\right) \\
&+ \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} - \frac{\gamma\eta IP}{4}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&+ \frac{\gamma^2\eta^2 LIP\rho^2\sigma^2}{2}
\end{aligned}
$$

where the last inequality is because $\gamma \leq \frac{1}{12LIP}$ thus $-(1 - 72\gamma^2 L^2 I^2 P^2) \leq -\frac{1}{2}$, and $\gamma\eta \leq \frac{1}{LIP}$ thus $\frac{1}{IP} - \gamma\eta L \geq 0$.

Rearranging, we have

$$
\begin{aligned}
&\frac{\gamma\eta IP}{4}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&\leq \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right] + \frac{9\gamma^3\eta L^2 I^2 P}{2}\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right) \\
&\quad + \frac{3\gamma\eta L^2 IP}{2}\left(3\gamma^2 I^2\tilde{\nu}^2 + 24\gamma^2 I^2 P^2\tilde{\beta}^2 + 3\gamma^2 IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]\right) \\
&\quad + \frac{3\gamma\eta IP\tilde{\delta}^2(P)}{2} + \frac{\gamma^2\eta^2 LIP\rho^2\sigma^2}{2}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&\leq \frac{4\left(\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right]\right)}{\gamma\eta IP} + 18\gamma^2 L^2 I\left(6I\tilde{\nu}^2 + \sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right) \\
&\quad + 6L^2\left(3\gamma^2 I^2\tilde{\nu}^2 + 24\gamma^2 I^2 P^2\tilde{\beta}^2 + 3\gamma^2 IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ 6\tilde{\delta}^2(P) + 2\gamma\eta L\rho^2\sigma^2 \\
&= \frac{4\left(\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right]\right)}{\gamma\eta IP} + 126\gamma^2L^2I^2\tilde{\nu}^2 + 144\gamma^2L^2I^2P^2\tilde{\beta}^2 + 6\tilde{\delta}^2(P) \\
&\quad + 18\gamma^2L^2I\sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right) + 18\gamma^2L^2IP\sigma^2\left[\rho^2 + \frac{1}{6P}\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right)\right] + 2\gamma\eta L\rho^2\sigma^2 \\
&= \frac{4\left(\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right]\right)}{\gamma\eta IP} + 126\gamma^2L^2I^2\tilde{\nu}^2 + 144\gamma^2L^2I^2P^2\tilde{\beta}^2 + 6\tilde{\delta}^2(P) \\
&\quad + 21\gamma^2L^2I\sigma^2\left(1 - \sum_{n=1}^{N}(q_t^n)^2\right) + 18\gamma^2L^2IP\sigma^2\rho^2 + 2\gamma\eta L\rho^2\sigma^2 \\
&\leq \frac{4\left(\mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0})\right] - \mathbb{E}_{\cdot|\mathcal{Q}}\left[f(\mathbf{x}_{t_0+P})\right]\right)}{\gamma\eta IP} + 126\gamma^2L^2I^2\tilde{\nu}^2 + 144\gamma^2L^2I^2P^2\tilde{\beta}^2 + 6\tilde{\delta}^2(P) \\
&\quad + \left(21\gamma^2L^2I + 18\gamma^2L^2IP\rho^2 + 2\gamma\eta L\rho^2\right)\sigma^2. \qquad \text{(C.16)}
\end{aligned}
$$

We now have

$$
\begin{aligned}
&\min_t \mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\
&\leq \min_{t_0 \in \left\{0, P, 2P, \ldots, \left(\lfloor \frac{T}{P} \rfloor - 1\right)P\right\}} \mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&\leq \frac{1}{\lfloor T/P \rfloor} \cdot \sum_{t_0 = 0, P, 2P, \ldots, \left(\lfloor \frac{T}{P} \rfloor - 1\right)P} \mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\nabla f(\mathbf{x}_{t_0})\|^2\right] \\
&\overset{(a)}{\leq} \frac{8\left(f(\mathbf{x}_0) - f^*\right)}{\gamma\eta IT} + 126\gamma^2L^2I^2\tilde{\nu}^2 + 144\gamma^2L^2I^2P^2\tilde{\beta}^2 + 6\tilde{\delta}^2(P) \\
&\quad + \left(21\gamma^2L^2I + 18\gamma^2L^2IP\rho^2 + 2\gamma\eta L\rho^2\right)\sigma^2 \\
&= \mathcal{O}\left(\frac{\mathcal{F}}{\gamma\eta IT} + \gamma^2L^2I^2\tilde{\nu}^2 + \gamma^2L^2I^2P^2\tilde{\beta}^2 + \tilde{\delta}^2(P) + \left(\gamma^2L^2I + \gamma^2L^2IP\rho^2 + \gamma\eta L\rho^2\right)\sigma^2\right)
\end{aligned}
$$

where the first term in (a) is from telescoping of (C.16) and $\frac{1}{\lfloor T/P \rfloor P} \leq \frac{1}{(T/P-1)P} = \frac{1}{T-P} \leq \frac{1}{T-T/2} = \frac{2}{T}$ since $P \leq \frac{T}{2}$, and the other terms in (a) are constants independent of t_0. □

C.4 Proof of Corollaries 3.2 and 3.3

For Corollary 3.2, we note that $\gamma\eta = \min\left\{\frac{\sqrt{\mathcal{F}}}{\sigma\rho\sqrt{LIT}}; \frac{1}{LIP}\right\}$ by the choice of γ and η. Because $\gamma \leq \frac{1}{12LIP}$ (since $T \geq 1$) and $\gamma\eta \leq \frac{1}{LIP}$, the conditions of Theorem 3.1 hold. Then, using $\gamma\eta \leq \frac{\sqrt{\mathcal{F}}}{\sigma\rho\sqrt{LIT}}$, $\frac{1}{\gamma\eta} = \max\left\{\frac{\sigma\rho\sqrt{LIT}}{\sqrt{\mathcal{F}}}; LIP\right\} \leq \frac{\sigma\rho\sqrt{LIT}}{\sqrt{\mathcal{F}}} + LIP$, and $\rho^2 \leq 1$, we obtain the result.

For Corollary 3.3, we have $\gamma\eta = \frac{\sqrt{\mathcal{F}}}{\rho\sqrt{LIT}}$ according to the choice of γ and η. Because $\frac{\sqrt{\mathcal{F}}}{\rho\sqrt{LIT}} \leq \frac{1}{LIP}$, we have $\gamma\eta \leq \frac{1}{LIP}$ and the conditions of Theorem 3.1 hold. The result is obtained by plugging the values of γ and η into the result in Theorem 3.1, and noting that $\rho^2 \leq 1$ in the last term. □

D Additional Details and Results of Experiments

D.1 Additional Details of the Setup

Datasets. The FashionMNIST dataset (`https://github.com/zalandoresearch/fashion-mnist`) has MIT license and a citation requirement [34]. The CIFAR-10 dataset (`https://www.cs.toronto.edu/~kriz/cifar.html`) only has a citation requirement [19]. We have cited both papers in our main paper. Our experiments used the original split of training and test

data in each dataset, where the training data is further split into workers in a non-IID manner according to the labels of data samples with a similarity of 5%, as described in the main paper.

Motivation of the Setup. The non-IID splitting of data to clients and the periodic connectivity are motivated by application scenarios of FL. An example of a real-world scenario is where different geographically-dispersed user groups can participate in training at different times of the day, e.g., when their phones are charging during the night. Another example is where multiple organizations have their local servers (also known as edge servers), which can participate in training only when their servers are idle, such as during the night when there are few user requests, and these servers are located in different geographical regions. Our notion of "client" is a general term that can stand for the phone or the local server in these specific scenarios.

Availability of Clients. In the case of *periodically available* clients, as described in the main paper, subsets of clients with different majority class labels take turns to become available. For example, only those clients with the first two majority classes of data are available in the first 100 rounds; then, in the next 100 rounds, only those clients with the next two majority classes of data are available, and so on. The majority class of each client is determined by the non-IID data partitioning. There are 10 classes in total, so a full cycle of participation of all clients is 500 rounds. To avoid inherent synchronicity across different simulation instances, we apply a random offset in the participation cycle at the beginning of the FL process. With this random offset, the number of rounds where the first subset of clients participate can be decreased to less than 100. For example, the first subset of clients may participate for 37 rounds (an arbitrary number that is less than or equal to 100), the second subset of clients then participates for the next 100 rounds, and so on. This random offset varies across different simulation instances. It represents the practical scenario where the FL process can start at any time in the day. In each round, $S = 10$ clients are selected to participate, according to a random permutation among the currently available clients. The selection of a small number of clients among the available clients can be due to resource efficiency considerations, for instance.

In addition to periodic availability, we also consider a setting where all the $N = 250$ clients are *always available*, and a subset of $S = 10$ clients are selected to participate in each round. Here, we consider two participation patterns. For independent participation, the subset of clients are selected randomly according to a uniform distribution, independently across rounds. For regularized participation, the clients are first randomly permuted and then S clients are selected sequentially from the permuted sequence of clients, so that all the N clients participate in FL once in every $\frac{N}{S}$ rounds.

Hyperparameter Choices. The initial local learning rates of $\gamma = 0.1$ and $\gamma = 0.05$ for FashionMNIST and CIFAR-10, respectively, were obtained from a learning rate search where the model was trained using $\gamma \in \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ while fixing $\eta = 1$. We considered the case of always available clients for this learning rate search. We chose the learning rate that gave the smallest training loss after $2,000$ rounds of training for FashionMNIST and $4,000$ rounds of training for CIFAR-10, where we observed that the best learning rate is the same for both independent and regularized participation with each dataset. The reason of considering always available clients for tuning the initial learning rates is because this is the ideal case which allows the model to make big steps in training, so it is reasonable to use these learning rates as starting points even in the case of periodic availability. The number of rounds that use the initial learning rate in the case of periodic availability was determined so that the model's performance no longer improves substantially afterwards, if keeping the same initial learning rate.

The subsequent learning rates for the methods shown in Figure 1, with periodically available clients, are found from a search on a grid that is $\{1, 0.1, 0.01, 0.001, 0.0001\}$ times the initial learning rate, as described in the main paper. The optimal learning rates found from this search are shown in Table D.1.

We fix the number of local updates to $I = 5$ and use a minibatch size of 16 for local SGD at each participating client. These parameters were not tuned specifically, since the relative difference between different methods will likely follow the same trend for different I and minibatch sizes. Because I is a controllable parameter in our experiments, we consider a minor extension (compared to our discussion in Section 5) in the "wait-*" methods in our experiments, where each client performs I local iterations (instead of one iteration in the discussion in Section 5) before starting to wait. The intuition remains the same as what has been discussed in Section 5. The reason for this extension is because the "wait-*" methods defined in this way is equivalent to the case where all the clients are available but the total number of rounds is reduced by P, so it has an easily interpretable meaning in

Table D.1: Optimal learning rates found from grid search

Method	FashionMNIST	CIFAR-10
wait-minibatch	$\gamma = 0.1$	$\gamma = 0.05$
wait-full	$\gamma = 0.1$	$\gamma = 0.05$
Algorithm 1 without amplification ($\eta = 1$)	$\gamma = 1 \times 10^{-5}$	$\gamma = 5 \times 10^{-5}$
Algorithm 1 with amplification ($\eta = 10, P = 500$)	$\gamma = 1 \times 10^{-5}$	$\gamma = 5 \times 10^{-6}$

practice. In addition, changing I can change the performance of all methods (including Algorithm 1, both with and without amplification, and the "wait-*" methods), while their relative difference can remain similar. We do not study the effect of I and minibatch size in details in this work.

Model Architecture. The CNN architectures used in our experiments are shown in Table D.2, where we use Kaiming initialization [15] for all the ReLU layers. All convolutional and dense layers, except for the last layer, use ReLU activation.

Table D.2: CNN architectures

FashionMNIST	CIFAR-10
Convolutional (input: 1, kernel: 5×5, padding: 2, output: 32)	Convolutional (input: 3, kernel: 5×5, padding: 2, output: 32)
MaxPool (kernel: 2×2, stride: 2)	MaxPool (kernel: 2×2, stride: 2)
Convolutional (input: 32, kernel: 5×5, padding: 2, output: 32)	Convolutional (input: 32, kernel: 5×5, padding: 2, output: 64)
MaxPool (kernel: 2×2, stride: 2)	MaxPool (kernel: 2×2, stride: 2)
Dense (input: $1,568$, output: 128)	Dense (input: $4,096$, output: 512)
Dense (input: 128, output: 10)	Dense (input: 512, output: 128)
Softmax	Dense (input: 128, output: 10)
	Softmax

Collecting Results. The plots in Figures 1–2 and all the figures showing additional results in the next section were obtained from simulations with 10 different random seeds for FashionMNIST and 5 different random seeds for CIFAR-10. For the best viewing experience, we further applied moving average over a window length of 3% of the data points (on the x-axis) in the loss and accuracy results. In all the figures, the solid lines and the shaded areas show the average and standard deviation values, respectively, over the moving average window and all simulation runs for each configuration.

Computational Resources. The workload of experiments was split between a desktop machine with RTX 3070 GPU and an internal GPU cluster. On RTX 3070, the FashionMNIST experiment with $300,000$ training rounds takes about a day to complete for a single simulation instance; the CIFAR-10 experiment with $600,000$ training rounds takes about three days to complete for a single simulation instance. The RTX 3070 has memory that can support simulation with 5 different random seeds (i.e., 5 different instances) at the same time. On the internal GPU cluster, the running times are slightly shorter since it has better-performing GPUs. We also note that the "wait-*" methods run faster than Algorithm 1 in simulation because a lot of rounds are skipped, but this does not mean that they can train models faster in practice, because we consider the number of training rounds as an indication of physical time elapse. For example, in the case of periodic availability, each subset of clients is available for 100 rounds in a cycle. This means that 100 rounds of training can take place during the physical time duration that these clients are available. Algorithm 1 chooses to actually train models during this time. The "wait-*" methods choose not to train in some of the rounds, but they still need to wait until the time has passed and the next subset of clients becomes available, where the time here is measured by the number of training rounds that includes those rounds that may be skipped by the "wait-*" methods.

D.2 Additional Results

In the following, we present some additional results obtained from experiments.

D.2.1 Always Available Clients

We first consider the case of always available clients, where a subset of $S = 10$ clients are selected in an independent or regularized manner, as discussed in Section D.1. We compare the results of independent and regularized participation in Figure D.1. The observation is that both methods give similar performance, with regularized participation having a marginal performance improvement. As our theory predicts (see Table 1), both independent and regularized participation can guarantee convergence, where regularized participation gives a slightly better convergence rate.

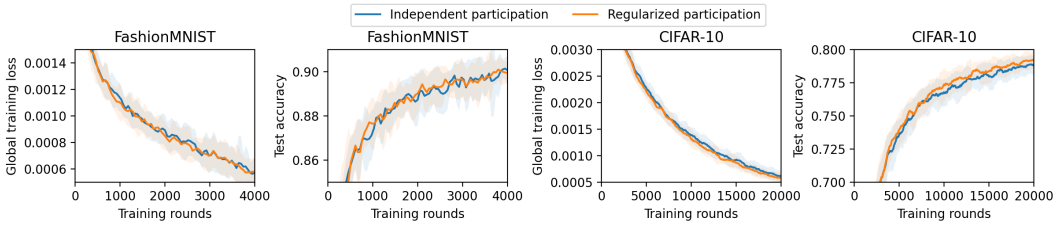


Figure D.1: Comparison between independent and regularized participation in the case of always available clients.

D.2.2 Periodically Available Clients

Different Learning Rates and Amplification Factors. Next, we consider periodically available clients as in the main paper. We study the performance of Algorithm 1 with different learning rate and amplification factor configurations. The results are shown in Figures D.2 and D.3.

Figure D.2 shows the results with different combinations of the local learning rate γ and amplification factor η. We note that $\eta = 1$ corresponds to the case without amplification and $\eta = 10$ corresponds to the case with amplification (and with an amplification factor of 10). We can observe that when γ is large, the training is generally unstable with high fluctuation, high loss, and low accuracy. Reducing γ improves the performance. According to our theory, $P = 500$ gives a small $\tilde{\delta}^2(P)$ with this experimental setup, because the cycle of all clients being available once is 500 rounds. On the other hand, a small $\tilde{\delta}^2(P)$ can be achieved with a much smaller P in the case of always available clients (in Section D.2.1). When P gets large, our theory suggests that the learning rate γ should become small in order to guarantee convergence. This explains why non-convergence is observed in Figure D.2 for large γ, while convergence is observed for small γ. For the smallest γ in each case, i.e., $\gamma = 10^{-5}$ for FashionMNIST and $\gamma = 5 \times 10^{-6}$ for CIFAR-10, we can see that amplification with $\eta = 10$ further improves the performance over no amplification ($\eta = 1$).

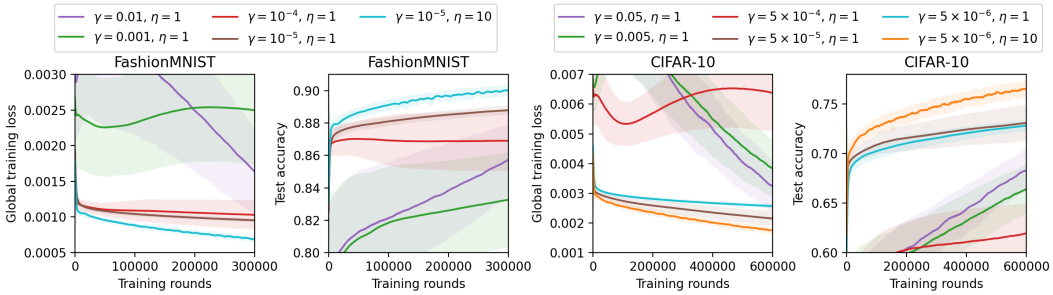


Figure D.2: Algorithm 1 with different learning rates in the case of periodically available clients. Note: FashionMNIST with $\gamma = 0.1$ is not shown in the plots due to NaN values observed when the number of training rounds is large.

In Figure D.3, we further investigate the impact of different amplification factors η, while fixing $\gamma = 10^{-5}$ for FashionMNIST and $\gamma = 5 \times 10^{-6}$ for CIFAR-10 as in the case of $\eta = 10$ in Figure D.2.

We observe that the cases with $\eta = 5$, $\eta = 10$, and $\eta = 15$ have similar performance, especially when getting closer to convergence. The performance gets worse when η is too small (e.g., $\eta = 2$), because the advantage of amplification is not fully attained in this case. When η is too large (e.g., $\eta = 20$), we observe higher fluctuation in the loss and accuracy values, which also leads to lower performance. For both datasets and their corresponding models considered in our experiments, we observe that the cases of $\eta = 5$, $\eta = 10$, and $\eta = 15$ all give reasonably good performance. This suggests that a coarse choice of η based on experience can be sufficient in practice, without the need of fine-grained tuning.

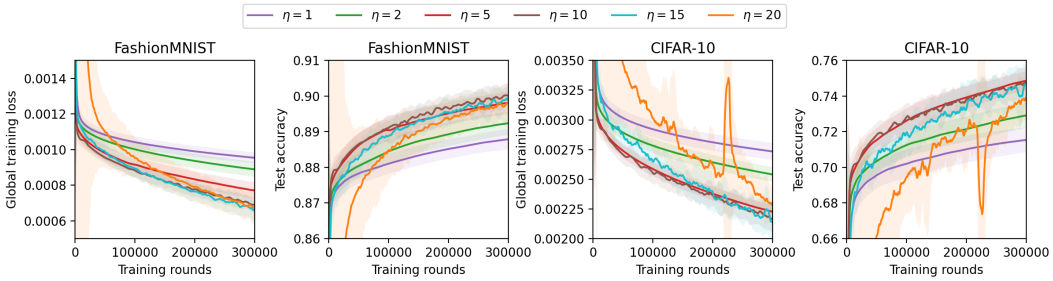


Figure D.3: Algorithm 1 with different amplification factors η, while fixing $\gamma = 10^{-5}$ for FashionMNIST and $\gamma = 5 \times 10^{-6}$ for CIFAR-10.