On the Convergence of Federated Averaging with Cyclic Client Participation

Yae Jee Cho [1] **Pranay Sharma** [1] **Gauri Joshi** [1] **Zheng Xu** [2] **Satyen Kale** [2] **Tong Zhang** [3 2]

Abstract

Federated Averaging (FedAvg) and its variants are the most popular optimization algorithms in federated learning (FL). Previous convergence analyses of FedAvg either assume full client participation or partial client participation where the clients can be uniformly sampled. However, in practical cross-device FL systems, only a subset of clients that satisfy local criteria such as battery status, network connectivity, and maximum participation frequency requirements (to ensure privacy) are available for training at a given time. As a result, client availability follows a *natural cyclic pattern*. We provide (to our knowledge) the first theoretical framework to analyze the convergence of FedAvg with cyclic client participation with several different client optimizers such as GD, SGD, and shuffled SGD. Our analysis discovers that cyclic client participation can achieve a faster asymptotic convergence rate than vanilla FedAvg with uniform client participation under suitable conditions, providing valuable insights into the design of client sampling protocols.

1. Introduction

Federated learning (FL) is a distributed learning framework that enables edge clients (e.g., mobile phones, tablets) to collaboratively train a machine learning (ML) model without sharing their local data (McMahan et al., 2017). In cross-device FL (Kairouz et al., 2019), millions of mobile devices are orchestrated by a central server for training, and only a subset of client devicess will participate in each communication round due to intermittent connectivity and resource constraints (Bonawitz et al., 2019).

Federated Averaging (FedAvg) (McMahan et al., 2017) and its variants (Reddi et al., 2021; Sahu et al., 2020; Wang et al., 2021) are the most popular algorithms in FL. In each communication round of the generalized FedAvg framework (Reddi et al., 2021; Wang et al., 2021): 1) the server broadcasts the current model to a subset of clients, 2) clients update the model with local data and send back the local model update, and 3) the server aggregates clients' model updates and computes the new global model. This algorithm is popular in practice for various reasons including the compatibility with FL system implementation (Bonawitz et al., 2019) and additional privacy techniques such as differential privacy (McMahan et al., 2018) and secure aggregation (Bonawitz et al., 2016).

The convergence of (generalized) FedAvg (also known as local SGD) has been studied in many recent works (Li et al., 2020b; Woodworth et al., 2020b; Wang et al., 2022; Karimireddy et al., 2019) due to its popularity in practice. While these analyses tackle the theoretical challenge of data heterogeneity, they assume either full client participation where all clients will participate every round, or partial client participation where the clients are chosen uniformly at random from the entire set of clients. However, in practical cross-device FL systems, clients can only participate in training when local criteria such as being idle, plugged in for charging, and on an unmetered network are satisfied (Bonawitz et al., 2019; Hard et al., 2018; Paulik et al., 2021; Huba et al., 2022). Works like Yang et al. (2018); Eichner et al. (2019); Zhu et al. (2021) observe client participation to have a diurnal pattern, and Balle et al. (2020); Kairouz et al. (2019); Wang et al. (2021) discuss the difficulty of controlling the sampling of clients for participation. Motivated by differential privacy (Kairouz et al., 2021a), McMahan & Thakurta (2022) seeks to limit the contribution of each client by allowing it to participate at most once in a large time window. The client participation control is then widely applied in training Gboard language models (Xu et al., 2023). Similarly, concurrent work has also looked into *regularized* client participation in FL and its convergence. For these reasons, clients typically participate in training with a *cyclic pattern* in practical FL systems.

In this work, we provide the first (to the best of our knowledge) convergence analysis of federated averaging with cyclic client participation. We consider that clients are implicitly divided into groups, and the groups become available to the server in a cyclic order. We show that for a global PL

[1]Carnegie Mellon University, USA [2]Google Research, USA [3]Hong Kong University of Science and Technology, Hong Kong. Correspondence to: Yae Jee Cho <yaejeec@andrew.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

objective (Haddadpour et al., 2019), instead of the standard $\mathcal{O}(1/T)$ rate of error convergence achieved by FedAvg, where T is the number of communication rounds, cyclic client participation can achieve a faster $\widetilde{\mathcal{O}}\left(1/T^2\right)$ convergence under suitable conditions, where $\widetilde{\mathcal{O}}(\cdot)$ subsumes all log-terms and constants. This key insight is similar to that obtained by a recent work (Yun et al., 2022) on the convergence of mini-batch and local-update shuffle SGD, which shows the fast convergence of local data shuffling at clients under the full (rather than cyclic and partial) client participation setting (see Section 2.2 for more details).

Our analysis framework covers several cases of cyclic participation and different client optimizers: 1) it includes the subsampling of a subset of clients from each group that becomes cyclically available, 2) it captures how the number of groups within a cycle or the data heterogeneity characteristics of the client groups affect convergence, and 3) it covers different client local procedures including gradient descent (GD), stochastic gradient descent (SGD), and shuffled SGD (SSGD). As a result of this generality, several well-studied FedAvg variants such as standard FedAvg with partial client participation (Li et al., 2020b; Jhunjhunwala et al., 2022) , minibatch RR and local RR (Yun et al., 2022) can become special cases of our framework. We show that our bounds match with the bounds from prior works in these special cases, corroborating the validity of our results. We also present preliminary experimental results to demonstrate that cyclic client participation indeed achieves better performance in terms of test accuracy and training loss convergence compared to standard FedAvg.

2. Related Work

2.1. Client Participation in FL.

Due to the large total number of clients in cross-device FL, it is inevitable to select only a subset of clients per training round. Therefore, there has been a plethora of work related to client participation in FL (Kairouz et al., 2019; Li et al., 2020a). Most work has focused on analyzing FedAvg with unbiased partial client participation (Yang et al., 2021; Jhunjhunwala et al., 2022) and showing a convergence rate of $\mathcal{O}(1/T)$. While some work in FL has also considered biased partial client participation for flexible client participation (Ruan et al., 2020) or loss-dependent client participation (Cho et al., 2020; Goetz et al., 2019), cyclic participation patterns have not been considered in these previous work.

Another related line of work is the analyses on arbitrary client participation presented in recent work (Wang & Ji, 2022; Avdyukhin & Kasiviswanathan, 2021). Wang & Ji (2022) proposes classes of different client participation patterns where cyclic client participation goes under the regularized participation class. However, due to the generality of the formulation, their analysis does not capture important characteristics such as how the ordering of the clients or the number of client groups within a cycle affects the convergence. Avdyukhin & Kasiviswanathan (2021) analyzes FedAvg with clients sending their local updates in an asynchronous manner, where each client has its own-defined cycle interval for sending its updates. However, such framework does not simulate the cyclic pattern that a realistic FL system observes where groups of clients sequentially become available to the server.

Cyclic client participation has only recently been viewed in FL through the lens of privacy (Kairouz et al., 2021b; Choquette-Choo et al., 2022) and communication-efficiency (Zhu et al., 2022). While Kairouz et al. (2021b) shows that cyclic client participation can improve privacy guarantees in FL, its convergence properties are not examined. Zhu et al. (2022) shows that selecting clients based on their participation frequencies can speed up convergence with the rate $\mathcal{O}(1/TV)$ where V is a constant depending on the variance arising from the data heterogeneity with partial client participation. However, the exact rate of the convergence speed-up is unclear due to the lack of bounds for the variable V. In contrast to this prior work, we provide the convergence for cyclic client participation in FL where the speed-up rate is clear (at the rate $\widetilde{\mathcal{O}}\left(1/T^2\right)$) and the conditions under which it can be achieved are identified. This speedup relies on analyzing FedAvg with cyclic participation from the perspective of shuffling-based methods which we explain in more detail below. A concurrent work (Malinovsky et al., 2023) has studied the convergence of regularized client participation where clients participate only once during a certain number of communication rounds, showing the best achievable rate of $\mathcal{O}\left(1/T^2\right)$ similar to our work. However, the corresponding work does not identify in which cyclic pattern the $\mathcal{O}\left(1/T^2\right)$ convergence can be achieved, and how intra-client and inter-client group data heterogeneity affects the convergence.

2.2. Shuffling-based methods.

The initial progress on shuffling-based methods was made by (Gurbuzbalaban et al., 2019; Gürbüzbalaban et al., 2021) for strongly-convex quadratics. The general idea in these, and subsequent works is that since shuffling-based methods involve using each component function *exactly once* in each epoch, the progress made by these methods within an epoch approximates that of full-batch gradient descent.

The literature on shuffling-based methods mainly focuses on three kinds of epochs: (i) random reshuffling (RR), where the data is shuffled after every epoch, (ii) shuffle once (SO), where the data is shuffled just once at the beginning, and (iii) incremental gradient (IG) method, in which the data is not

Figure 1. Illustration of cyclic client participation (CyCP) with $M = 12$ clients divided into $\overline{K} = 3$ groups. In each communication round, $N = 2$ clients are selected for training from the client group available at that time. All groups are traversed once in a *cycle-epoch* consisting of $\overline{K}$ communication rounds.

shuffled at all, and follows a predetermined order in each epoch. We shall see in Section 3 (and more so in Theorem 1) that cyclic client participation essentially approximates an incremental gradient method at the level of the server, with each client interpreted as a sample.

Recent work has established upper and lower bounds for shuffle SGD under these shuffling schemes. For RR (and SO), (Safran & Shamir, 2020) showed a lower bound of $\mathcal{O}(\frac{1}{n^2K^2} + \frac{1}{K^3})$ for strongly convex quadratic (K is the number of epochs, n is the number of samples), while (Rajput et al., 2020) showed $\mathcal{O}(\frac{1}{nK^2})$ lower bound for general strongly-convex F with smooth $\{f_i\}$. Matching upper bounds have been achieved in the large epoch regime by (Ahn et al., 2020; Mishchenko et al., 2020) for smooth PL functions and (Nagaraj et al., 2019) for smooth strongly convex functions. For IG, (Nguyen et al., 2021) showed $\mathcal{O}(1/K^2)$ rate for strongly-convex F with smooth $\{f_i\}$. The improved dependence on K in all these works requires K to be larger than $\mathcal{O}(\kappa^a)$, where κ is the condition number of the problem, and $a \in [1, 2]$. This large epoch requirement has been shown to be essential in (Safran & Shamir, 2021).

3. Problem Formulation

System Model and Objectives. Consider a cross-device FL setting where we have M total clients. Each client $m \in [M]$ has its local training dataset $\mathcal{B}_m$ and its corresponding local empirical loss function $F_m(\mathbf{w}) = \frac{1}{|\mathcal{B}_m|}\sum_{\xi\in\mathcal{B}_m} \ell(\mathbf{w}, \xi)$, where $\ell(\mathbf{w}, \xi)$ is the loss value for the model $\mathbf{w} \in \mathbb{R}^d$ at data sample ξ. The optimization task is identical to that of standard FL (McMahan et al., 2017; Kairouz et al., 2019) where the global objective is $F(\mathbf{w}) = \frac{1}{M}\sum_{m=1}^{M} F_m(\mathbf{w})$ and the server aims to find the model that achieves $\min_{\mathbf{w}} F(\mathbf{w})$. Throughout the paper, all vector and matrix norms are Euclidean and spectral norms, respectively.

Cyclic Client Participation (CyCP). We consider that the M clients are divided into $\overline{K}$ non-overlapping client groups such that each group contains $M/\overline{K}$ clients, as illustrated in Fig. 1. The client groups are denoted by $\sigma(i),\ i \in [\overline{K}]$, where each $\sigma(i)$ contains the associated clients' indices. The groups and the order in which they are traversed by the server (say, $\sigma(1), \ldots, \sigma(\overline{K})$) are pre-determined and fixed throughout training to simulate a cyclic structure of client participation. In each communication round, once a client group $\sigma(i)$ becomes available, the server selects a subset of N clients from $\sigma(i)$ uniformly at random without replacement. As a result of the cyclic structure, and subsampling within each available client group, once selected, a client cannot participate in training at least for the next $\overline{K} - 1$ rounds. For brevity, we call this cyclic client participation framework as CyCP throughout the paper.

Observe that the CyCP framework reflects several practical FL scenarios mentioned in Section 1. Each client can participate at most once in $\overline{K}$ consecutive communication rounds in CyCP, which satisfies the privacy requirements (Kairouz et al., 2021b; Choquette-Choo et al., 2022). A timer on each client can be used to enforce that clients can only participate in training again after a certain period, and this period corresponds to $\overline{K}$. Even without enforcing the timer-based criterion, CyCP captures the natural participation pattern due to clients coming from different time zones or preference of charging their devices (Paulik et al., 2021; Huba et al., 2022; Yang et al., 2018).

We introduce the term "cycle-epoch" to refer to the interval in which the server goes through all the client groups $\{\sigma(i),\ i \in [\overline{K}]\}$ once sequentially. In other words, each cycle-epoch consists of $\overline{K}$ communication rounds. Formally, we set k as the index for the cycle-epoch, and $i \in [\overline{K}]$ as the index for the currently available client group within a cycle-epoch. The server sends the global model $\mathbf{w}^{(k,i-1)}$ to the set $\mathcal{S}^{(k,i)}$ of N clients, selected from the client group $\sigma(i)$, to perform local training. We consider three different types of client local updates which we explain in detail below.

Client Local Update. Each client $m \in \mathcal{S}^{(k,i)}$ initializes its local model as $\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$ and performs local update(s). The global model is updated as:

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} + \Delta^{(k,i-1)} \tag{1}$$

where $\Delta^{(k,i-1)}$ is the aggregate of local updates from clients in $\mathcal{S}^{(k,i)}$. We consider three different client local update procedures for our CyCP framework.

(i) Local Gradient Descent (GD): Selected clients perform a single GD step to update their local model. Therefore,

$$\Delta^{(k,i-1)} = -\frac{\eta}{N}\sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)})$$

The resulting global algorithm is referred to as FedSGD in (McMahan et al., 2017).

(ii) Local Stochastic Gradient Descent (Local SGD): To avoid the cost of computing full gradients, each client performs τ local updates to its model using stochastic gradients $\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$ computed using

a minibatch $\xi_m^{(k,i-1,l)}$ sampled uniformly at random from client m's local dataset $\mathcal{B}_m$. Thus, the client's model update is

$$\Delta^{(k,i-1)} = -\frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$$

with $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$.

(iii) Local Shuffled SGD (SSGD): Recent works in FL (Yun et al., 2022; Malinovsky et al., 2021) propose the use of local SSGD, where clients partition their local datasets into B disjoint *components*, that is, the local loss at client m can be expressed as $F_m(\mathbf{w}) = \frac{1}{B}\sum_{l=0}^{B-1} F_{m,l}(\mathbf{w})$. We define $\mathcal{P}_B$ to be the set of all permutations of $\{0, ..., B-1\}$. In each round, the client performs local updates by going over all the components, in an order decided by the random permutation $\pi_m^k \sim \text{Unif}(\mathcal{P}_B)$. The resulting model update is

$$\Delta^{(k,i-1)} = -\frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)})$$

with $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)})$.

Further details of our framework of FL with CyCP are shown in Algorithm 1.

Special Cases of CyCP. The CyCP framework covers different algorithms such as standard FedAvg with partial client participation or minibatch RR and local RR presented in (Yun et al., 2022). When $\overline{K} = 1$, the CyCP setting becomes *standard FedAvg* with partial client participation where in each round, N clients are sampled from the same entire client population. When $\overline{K} = 1$ and $N = M$, CyCP with local GD becomes identical to minibatch RR (Yun et al., 2022) with M clients, each with a single component. Both converge exponentially fast to the optimum. Another special case is when we have $\overline{K} = 1$ with $N = M$ but for local SSGD, in which case we have local RR (Yun et al., 2022) with B components at each client, with synchronization of the updates happening for every B components. We show in Section 4 that our theoretical results match the bounds accordingly for these special cases.

4. Convergence Analysis

In this section, we provide and compare the convergence bounds for CyCP in FL for the three client local update methods described above, and provide insights into how the achieved complexities with CyCP ($\overline{K} > 1$) compare with standard FedAvg ($\overline{K} = 1$). All the proofs are deferred to Appendix A-C.

Algorithm 1 CyCP Framework in FL

Input: Global Model $\mathbf{w}^{(1,0)}$, Client groups $\sigma(i), i \in [\overline{K}]$
Output: Global Model $\mathbf{w}^{K+1,0}$
For $k \in [K]$ cycle-epochs do: # *Cyclic Participation*
  For $i \in [\overline{K}]$ do: # $T = K\overline{K}$ *comm. rounds*
    Sample N clients from client set $\sigma(i)$ uniformly at random w/o replacement to get client set $\mathcal{S}^{(k,i)}$.
    Send global model $\mathbf{w}^{(k,i-1)}$ to clients in $\mathcal{S}^{(k,i)}$.
    Clients $m \in \mathcal{S}^{(k,i)}$ in parallel do:
      $\mathbf{w}_m^{(k+1,0)} \leftarrow$ **LocalUpdate**$(m,\ \mathbf{w}^{(k,i-1)}, case)$
    $\mathbf{w}^{(k,i)} = \frac{1}{N}\sum_{m \in \mathcal{S}^{(k,i)}} \mathbf{w}_m^{(k+1,0)}$
  $\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,\overline{K})}$

LocalUpdate$(m,\ \mathbf{w},\ case)$:
  Set local model $\mathbf{w}_m = \mathbf{w}$
  if $case == LocalGD$:
    Update $\mathbf{w}_m \leftarrow \mathbf{w}_m - \eta \nabla F_m(\mathbf{w}_m)$
  elif $case == LocalSGD$:
    For $j \in [\tau]$ do:
      Sample mini-batch ξ from local dataset $\mathcal{B}_m$
      Update $\mathbf{w}_m \leftarrow \mathbf{w}_m - \eta \nabla F_m(\mathbf{w}_m,\ \xi_m)$
  elif $case == ShuffledSGD$:
    Sample $\pi_m^k \sim \text{Unif}(\mathcal{P}_B)$
    For $j \in [B]$ do:
      Update $\mathbf{w}_m \leftarrow \mathbf{w}_m - \eta \nabla F_{m,\pi_m^k(j-1)}(\mathbf{w}_m)$

4.1. Assumptions

First, we present the assumptions used for the convergence guarantees in this work.

Assumption 1 (Smoothness of $F_m(\mathbf{w}),\ \forall\, m$)**.** The clients' local objective functions $F_1(\mathbf{w}),\ ..., F_M(\mathbf{w})$, are all L-smooth, that is, $\|\nabla F_m(\mathbf{w}) - \nabla F_m(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$ for all m, $\mathbf{w}$ and $\mathbf{w}'$.

Assumption 2 (μ-Polyak-Łojasiewicz $F(\mathbf{w})$)**.** For some $\mu > 0$, the global objective satisfies $\frac{1}{2}\|\nabla F(\mathbf{w})\|^2 \geq \mu(F(\mathbf{w}) - \min_{\mathbf{w}'} F(\mathbf{w}'))$ for all $\mathbf{w}$.

Assumption 1, 2 are common in the optimization and FL literature (Haddadpour & Mahdavi, 2019; Karimi et al., 2020; Haddadpour et al., 2019; Gower et al., 2021). While we restrict ourselves to PL functions for brevity and for ease of comparison with prior work (Yun et al., 2022), the analyses can be generalized to general nonconvex functions using techniques proposed in (Li et al., 2021).

Next, we present the assumptions over the client groups $\sigma(1), \ldots, \sigma(\overline{K})$.

Assumption 3 (Intra-group & Inter-Group Data Heterogeneity)**.** There exist constants $\gamma,\ \alpha \geq 0$, such that for all $\mathbf{w}$, for all $i \in [\overline{K}]$ and for all $m \in \sigma(i)$,

$\|\nabla F_m(\mathbf{w}) - \frac{1}{|\sigma(i)|}\sum_{m\in\sigma(i)} \nabla F_m(\mathbf{w})\| \leq \gamma$, and
$\|\frac{1}{|\sigma(i)|}\sum_{m\in\sigma(i)} \nabla F_m(\mathbf{w}) - \nabla F(\mathbf{w})\| \leq \alpha$.

Assumption 3 bounds the data heterogeneity across clients *within a group* by γ and the data heterogeneity *across groups* by α. Assumption 3 also implies the commonly used data heterogeneity assumption used in previous FL literature (Yu et al., 2019; Koloskova et al., 2020; Wang et al., 2020a;b; Reddi et al., 2021) as follows:

Lemma 1. *If Assumption 3 is true, there exists $\nu = \gamma + \alpha \geq 0$ such that $\left\|\nabla F_m(\mathbf{w}) - \frac{1}{M}\sum_{i=1}^{M} \nabla F_i(\mathbf{w})\right\| \leq \nu$ for all clients $m \in [M]$, and for all $\mathbf{w}$.*

Using Assumption 3 instead of the standard assumption allows us to derive tighter convergence bounds in terms of γ and α, and separate the effect of the two kinds of heterogeneity, as we discuss in subsequent sections.

4.2. Convergence for CyCP with Local GD

First, we start with providing the convergence of the global model in CyCP with local GD.

Theorem 1 (Convergence with CyCP+GD)**.** *With Assumptions 1, 2, 3, the choice of step-size $\eta = \log(MT^2/\overline{K}^2)/\mu NT$, and number of communication rounds $T \geq 7\kappa\overline{K}\log\left(MT^2/\overline{K}^2\right)$ where $\kappa = L/\mu$:*

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right), \tag{2}$$

where $\tilde{\mathcal{O}}(\cdot)$ subsumes all log-terms and constants.

Although it might appear that the bound becomes worse with increasing $\overline{K}$ due to it appearing in the numerators of the terms, since $T = K\overline{K}$ (see Algorithm 1), a large $\overline{K}$ has no adverse impact on the convergence.

Convergence Dependence on γ^2 and α. Theorem 1 shows that CyCP+GD converges at the rate of $\tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)$ which depends on γ^2, the intra-group data heterogeneity. Consequently, a large intra-group data heterogeneity γ leads to worse convergence. Conversely, if $\gamma \simeq 0$, CyCP+GD can achieve $\tilde{\mathcal{O}}\left(1/T^2\right)$ convergence, due to the $\tilde{\mathcal{O}}\left(1/T\right)$ domintnat term becoming zero. Hence, in the CyCP settings where clients within the same group have similar data distributions (i.e., γ is close to 0), CyCP+GD can yield a faster convergence rate compared to standard FedAvg ($\overline{K} = 1$). An example of a setting where this can naturally occur in realistic FL scenarios is when the cyclic patterns follow the diurnal-nocturnal pattern, also shown in (Zhu et al., 2021). It is also worth noting that the term with inter-group data heterogeneity α in Theorem 1 decays at the rate of $\tilde{\mathcal{O}}\left(1/T^2\right)$. Therefore, in CyCP settings, the intra-group data heterogeneity γ has a more significant contribution to the convergence error than the inter-group data heterogeneity α.

Convergence Dependence on $\overline{K}$. Theorem 1 also shows that even for $\gamma \neq 0$, CyCP+GD can gain a $\tilde{\mathcal{O}}\left(1/T^2\right)$ convergence rate when $\overline{K} = M/N$. This is a faster rate than the standard FedAvg (the setting with $\overline{K} = 1$) which has $\tilde{\mathcal{O}}\left(1/T\right)$ rate. While the convergence rates for the cases of $\overline{K} = 1$ and $\overline{K} = M/N$ are clear from Theorem 1, it is yet unclear what happens in the middle regime of $1 < \overline{K} < M/N$. For this, we compare the total cost of CyCP+GD and standard FedAvg to achieve ϵ error. We define the total communication and computation cost in one communication round of GD (which involves computing N gradients and communicating N vectors to the server) as c_{GD}. Then taking into account only the dominant term in (2), the total cost C_{GD} to achieve an ϵ error is

$$C_{\text{GD}}(\epsilon) = \tilde{\mathcal{O}}\left(\frac{c_{\text{GD}}\overline{K}\gamma^2}{\epsilon N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) \tag{3}$$

We compare $C_{\text{GD}}(\epsilon)$ with $\overline{K} > 1$ and $\overline{K} = 1$ denoted as $C_{\text{GD}|\overline{K}>1}(\epsilon)$, $C_{\text{GD}|\overline{K}=1}(\epsilon)$ respectively and derive the following result:

Corollary 1. *For the total cost defined as (3), for $\overline{K} < M/N$, we have that $C_{GD|\overline{K}>1}(\epsilon) > C_{GD|\overline{K}=1}(\epsilon)$.*

Corollary 1 shows that with the number of groups set to the middle range, i.e., $1 < \overline{K} < M/N$, CyCP does not incur a smaller cost compared to standard FedAvg ($\overline{K} = 1$). Hence, for CyCP+GD to incur a lower cost compared to standard FedAvg, the *necessary condition* is having $\overline{K} = M/N$. There may be some scenarios in which $\overline{K}$ is a naturally occurring quantity that the server does not have control over. It is worth noting that in these cases, N which is the number of selected clients per round, can be chosen accordingly by the server to pay a lower cost than standard FedAvg.

Matching Bounds with Minibath RR (Yun et al., 2022). Recall that with $\overline{K} = 1$ and $N = M$, CyCP+GD in Algorithm 1 becomes analogous to the minibatch RR algorithm with M clients where each client has a single component. In this case, our bound in Theorem 1 is only left with the first term that decays with the rate $\tilde{\mathcal{O}}\left(1/MT^2\right)$ which exactly matches minibatch RR's bound in [Theorem 1] (Yun et al., 2022) which shows exponential convergence.

4.3. Convergence for CyCP with Local SGD

Next, we present the convergence for CyCP with local SGD. Local SGD introduces additional technical challenges compared to GD for deriving the convergence analysis and re-

quires the following additional assumption over the stochastic gradients:

Assumption 4 (Bounded Variance)**.** For local objective $F_m(\mathbf{w})$, the local stochastic gradient $\nabla F_m(\mathbf{w}, \xi_m)$ computed using a mini-batch ξ_m, sampled uniformly at random from $\mathcal{B}_m$, has bounded variance, that is, $\mathbb{E}[\|\nabla F_m(\mathbf{w}, \xi_m) - \nabla F_m(\mathbf{w})\|^2] \leq \sigma^2$, for all $m \in [M]$.

Assumption 4 is commonly used in the stochastic optimization literature (Stich, 2019; Basu et al., 2019; Li et al., 2020b; Ruan et al., 2020). Now we present the convergence bound for local SGD.

Theorem 2 (Convergence with CyCP+SGD)**.** *With Assumptions 1, 2, 3, and 4 and step-size* $\eta = \log(MT^2/\overline{K}^2)/\tau\mu NT$*, for* $T \geq 10\kappa\overline{K}\log(MT^2/\overline{K}^2)$ *communication rounds where* $\kappa = L/\mu$*, the convergence error is bounded as:*

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{K}(\overline{K}-1)\alpha^2}{\mu T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\sigma^2}{\mu\tau NT}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2T^2}\right) \tag{4}$$

where $\tilde{\mathcal{O}}(\cdot)$ *subsumes all log-terms and constants.*

Again, although it might appear that the bound becomes worse with increasing $\overline{K}$, since $T = K\overline{K}$ in Algorithm 1, a large $\overline{K}$ has no adverse impact on the convergence.

Convergence Dependence on γ^2 and σ^2. In Theorem 2, the dominant $\widetilde{\mathcal{O}}(1/T)$ terms are dependent on two factors: the intra-group data heterogeneity γ^2 (which also appeared for local GD) and the stochastic gradient variance σ^2. Due to this, for $\sigma > 0$, even with $\overline{K} = M/N$ or $\gamma^2 \simeq 0$, we do not achieve the $\widetilde{\mathcal{O}}(1/T^2)$ convergence rate, as we did in the local GD case. Hence, even with CyCP, the best we can achieve when using local SGD is the convergence rate of $\widetilde{\mathcal{O}}(1/T)$. Seeing this result, one might wonder if there is any advantage at all of performing Local SGD with CyCP. We answer this question below by comparing the cost of CyCP with Local SGD to that of standard FedAvg.

Does CyCP ($\overline{K} > 1$) with Local SGD Ever Help for FL? At first glance of Theorem 2 one may think that CyCP does not improve the convergence rate with local SGD case due to stochastic gradient variance appearing in one of the dominant terms $\widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\sigma^2}{\mu\tau NT}\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)$. However, we show in Corollary 2 that this is not always the case. Similar to how we defined c_{GD} in the previous section, we definte the total communication and computation cost in one communication round with Local SGD as c_{SGD}. Formally, taking into account only the dominant terms in (4), the total cost to achieve an ϵ error for the local SGD case is:

$$C_{\text{SGD}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\text{SGD}}\overline{K}\gamma^2}{\epsilon N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) + \widetilde{\mathcal{O}}\left(\frac{c_{\text{SGD}}\sigma^2\overline{K}}{\epsilon N\tau}\right) \tag{5}$$

We denote the costs for CyCP and standard FedAvg as $C_{\text{SGD}|\overline{K}=1}$, $C_{\text{SGD}|\overline{K}>1}$ respectively. Now we show the conditions to have $C_{\text{SGD}|\overline{K}>1} < C_{\text{SGD}|\overline{K}=1}$, i.e., have CyCP incur a lower cost than standard FedAvg.

Corollary 2. *Suppose we have* $N = M/\overline{K}$*, and the intra-group data heterogeneity* γ *satisfies* $\gamma^2 \geq M\sigma^2/N\tau$*. Then, we get* $C_{SGD|\overline{K}} \leq C_{SGD|\overline{K}=1}$*.*

Corollary 2 shows that CyCP +SGD can indeed incur a lower cost to achieve ϵ error compared to standard FedAvg ($\overline{K} = 1$) when the intra-group data heterogeneity is sufficiently larger than the stochastic gradient variance divided by the number of local iterations. Note that the condition $\gamma^2 \geq M\sigma^2/N\tau$ in Corollary 2 can be satisfied by increasing the minibatch size b (which decreases the variance σ^2) or increasing the number of local iterations τ.

Matching Bounds with Standard FedAvg. For $\overline{K} = 1$, CyCP+SGD recovers Standard FedAvg with Local SGD, and our bound in (4) with full client participation follows the order of $\widetilde{\mathcal{O}}(\kappa^2\nu^2/\mu MT^2) + \widetilde{\mathcal{O}}(\kappa\sigma^2/\mu\tau MT)$. We show that this bound matches the last iterate bound in (Qu et al., 2020) which assumes full client participation, and bounded norm of the stochastic gradient with parameter G that reads $\widetilde{\mathcal{O}}(\kappa^2G^2/\mu T^2) + \widetilde{\mathcal{O}}(\kappa\sigma^2/\mu\tau T)$ where their learning rate doesn't decay with M as our case, leading to the lack of the $1/M$ in their bounds. The difference in the first term is due to their work assuming the bounded norm of the stochastic gradient G while we assume only the bounded variance of the stochastic gradient.

4.4. Convergence for CyCP with Local SSGD

For the last scenario of CyCP, we present results on the convergence properties of the global model with local SSGD. We slightly modify the local loss definition of each client as $F_m(\mathbf{w}) = \frac{1}{B}\sum_{l=0}^{B-1} F_{m,l}(\mathbf{w})$ so that each client has B loss components. For local SSGD, clients perform local updates sequentially over $F_{m,\pi_m(l)}(\mathbf{w})$, $l \in [0, ..., B-1]$ where $\pi_m \sim \text{Unif}(\mathcal{P}_B)$ is a random permutation over the B components and $\pi_m(l)$ denotes the l-th element of this permutation. For SSGD, in lieu of Assumption 4, we need the following intra-client component heterogeneity assumption, which is commonly used in the shuffled SGD literature (Yun et al., 2022; Malinovsky et al., 2021):

Assumption 5 (Intra-Client Component Heterogeneity)**.** There exists a constant $\overline{\nu} \geq 0$ such that for each client $m \in [M]$, and each component $l \in [0, ..., B-1]$ of its local dataset, $\|\nabla F_{m,l}(\mathbf{w}) - \nabla F_m(\mathbf{w})\| \leq \overline{\nu}$, for all $\mathbf{w}$.

Now we present our convergence results for the SSGD case.

Theorem 3 (Convergence with CyCP+SSGD)**.** *With Assumptions 1, 2, 3, and 5, and* $\eta = \log(MBT^2/\overline{K}^2)/\mu BT$ *and cycle-epoch* $T \geq 10\kappa\overline{K}\log\left(MBT^2/\overline{K}^2\right)$ *where* $\kappa = L/\mu$*, with probability at least* $1-\delta$*, the convergence error is bounded as:*

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* &\leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MBT^2} \\ &+ \widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) \\ &+ \widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{\nu}^2}{\mu T^2}\left(\frac{(B^{3/2}-1)^2}{B^4} + \frac{(B-1)^2}{B^3}\right)\right) \\ &+ \widetilde{\mathcal{O}}\left(\frac{\kappa\overline{K}\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)\end{aligned} \tag{6}$$

where $\tilde{\mathcal{O}}(\cdot)$ *subsumes all log-terms and constants.*

Again, since $T = K\overline{K}$, increasing $\overline{K}$ does not impact the bound above adversely.

Dependency on $\overline{K}$ is Identical to CyCP+GD. Theorem 3 shows that the dominant term is $\widetilde{\mathcal{O}}\left(\frac{\kappa\overline{K}\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)$ which is the same dominant term for the local GD's convergence rate in Theorem 1. This dominant term exists as long $\gamma \neq 0$ or $\overline{K} \neq M/N$. Hence, for $1 < \overline{K} < M/N$, as with local GD (see Corollary 1), CyCP +SSGD does not yield a lower cost than standard FedAvg. Also like CyCP+GD, $\overline{K} = M/N$ is *necessary* for CyCP+SSGD to get any cost reduction compared to standard FedAvg. We give a more detailed comparison between the costs of the two different local update procedures below.

Can CyCP+SSGD be better than CyCP+GD? We have seen above that CyCP+SSGD and CyCP+GD converge at the same rate due to the same dominant term which is non-zero for $\gamma^2 \neq 0$ and $\overline{K} < M/N$. For $\overline{K} = M/N$, however, the dominant term goes to 0 and CyCP+SSGD and CyCP+GD become comparable. Again, we define the total communication and computation cost in one communication round with local SSGD as c_{SSGD}. Then, for $\overline{K} = M/N$, we have that the total cost for CyCP+GD and CyCP+SSGD to achieve an ϵ error denoted as $C_{\text{GD}|\overline{K}=M/N}(\epsilon)$, $C_{\text{SSGD}|\overline{K}=M/N}(\epsilon)$ respectively, is

$$C_{\text{GD}|\overline{K}=M/N}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\text{SSGD}}}{\sqrt{\epsilon}}\left(\frac{\overline{K}}{\sqrt{M}} + \overline{K}\alpha\right)\right) \tag{7}$$

$$\begin{aligned}&C_{\text{SSGD}|\overline{K}=M/N}(\epsilon) \\ &= \widetilde{\mathcal{O}}\left(\frac{c_{\text{SSGD}}}{\sqrt{\epsilon}}\left(\frac{\overline{K}}{\sqrt{MB}} + \nu + \overline{K}\alpha + \frac{\overline{\nu}}{\sqrt{B}}\right)\right)\end{aligned} \tag{8}$$

With these costs, we have that CyCP+SSGD only incurs a lower cost than the CyCP+GD case for $\overline{K} = M/N$ when

$$1 - \frac{1}{\sqrt{B}} - \nu - \frac{\overline{\nu}}{\sqrt{B}} > 0,\ B > 1 \tag{9}$$

The only certain condition in which (9) can be satisfied is when $\nu \simeq \overline{\nu} \simeq 0$. Hence, with $\overline{K} = M/N$, only when there is close to 0 data heterogeneity across clients and intra-client component heterogeneity within each client is when CyCP+SSGD can incur a lower cost compared to CyCP+GD. This also aligns well with the theoretical results presented in (Yun et al., 2022; Woodworth et al., 2020a).

Matching Bounds with Special Cases of CyCP+SSGD. Special cases of CyCP+SSGD become analogous to different algorithms such as CyCP+GD or local RR (Yun et al., 2022). With $B = 1$, since $\overline{\nu} = 0$, we recover CyCP+GD (Theorem 1). Another special case is when $\overline{K} = 1,\ N = M$ for CyCP+SSGD. In this case, for each communication round, we have full client participation where each client performs SSGD over its local components, which becomes analogous to the local RR algorithm proposed and theoretically analyzed in (Yun et al., 2022). Note that in this case, the bound in Theorem 3 matches the bound for local RR presented in (Yun et al., 2022). We present a more detailed comparison in the following paragraph.

Is CyCP+SSGD better than Local RR? Since local RR assumes full client participation, for a fair comparison, we compare local RR with CyCP+SSGD with $\overline{K} = M/N$. The dominant term in CyCP+SSGD's convergence bound in Theorem 3 becomes zero and we are left with the terms having a convergence rate of $\widetilde{\mathcal{O}}\left(1/T^2\right)$. We compare the two algorithms in terms of the total cost to achieve ϵ error, where the total cost for local RR is:

$$C_{\text{LocalRR}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{\overline{K}c_{\text{SSGD}}}{\sqrt{\epsilon}}\left(\frac{1}{\sqrt{MB}} + \nu + \frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{10}$$

Note that for local RR, both the computation and communication cost is $\overline{K}$ times that of CyCP+SSGD, since in local RR all clients' participate in each communication round while in CyCP+SSGD only $M/\overline{K}$ clients participate per communication round. With CyCP+SSGD's cost in (8) for $\overline{K} = M/N$ and local RR's cost in (10), we have the following theoretical result that compares the two methods:

Corollary 3. *For a sufficiently large* M *such that*

$$M > N\left(1 + \frac{\alpha}{\gamma + \frac{\overline{\nu}}{\sqrt{B}}}\right) \tag{11}$$

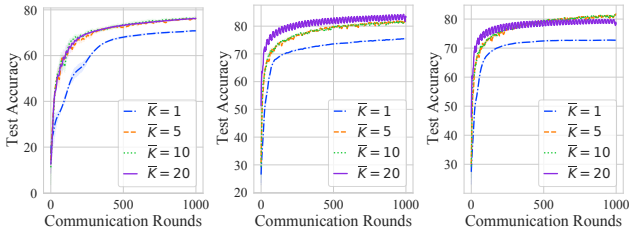
where $\overline{K} = M/N$ *for CyCP+SSGD, CyCP+SSGD is* ***always*** *better than local RR in terms of the total cost taken to gain epsilon error.*

Corollary 3 shows that with $\overline{K} = M/N$, and sufficiently large M, CyCP+SSGD is always preferred over local RR to achieve a lower cost. Since $\overline{K} = M/N$, a larger M indicates a larger $\overline{K}$. Observe that the lower bound on M in Corollary 3 becomes smaller for a smaller α and a larger γ. This indicates that it is preferred that the client groups have smaller inter-group data heterogeneity but larger intra-group data heterogeneity, for CyCP+SSGD to beat local RR.

5. Experimental Results

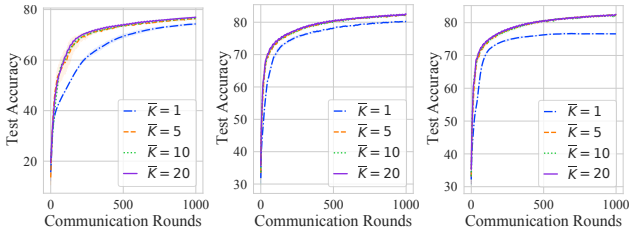
Setup. We train ML models on standard datasets using FedAvg with CyCP for different client local updated procedures to see how cyclicity affects the performance of FL. We experiment with image classification using an MLP for the FMNIST (Xiao et al., 2017) dataset and EMNIST dataset (Cohen et al., 2017) with 62 labels where we have 100 and 500 clients in total and select 5 and 10 clients per communication round respectively. We use the Dirichlet distribution $\text{Dir}_K(\alpha)$ (Hsu et al., 2019) to partition the data across clients where α determines the degree of the data heterogeneity across clients. Smaller α indicates larger data heterogeneity. We experiment with three different seeds for the randomness in the dataset partition across clients and present the averaged results. All experiments are conducted on clusters equipped with one NVIDIA TitanX GPU. The algorithms are implemented in PyTorch 1. 11. 0. The code used for all experiments is included in the supplementary material. For all experiments, we do a grid search over the required hyperparameters to find the best-performing ones and then fix the hyperparameters and only change $\overline{K}$. Specifically, we do a grid search over the learning rate: $\eta \in \{0.05, 0.01, 0.005, 0.001\}$, batch size: $b \in \{32, 64, 128\}$, and local iterations: $\tau \in \{5, 10, 30, 50\}$ to find the hyper-parameters with the highest test accuracy for each benchmark. For the deep multi-layer perceptron used for our experiments, we use a network with 2 hidden layers of units $[64, 30]$ with dropout after the first hidden layer where the input is the normalized flattened image and the output consists of the label space. For all experiments, the data is partitioned to $80\%/10\%/10\%$ for training/validation/test data, where the training data then is again partitioned across the clients heterogeneously.

Effect of $\overline{K}$ and Data Heterogeneity. We show in Fig. 2 the test accuracy for the FMNIST dataset for different $\overline{K}$ values and client local procedures for high data heterogeneity ($\alpha = 0.5$). Recall that $\overline{K} = 1$ is analogous to standard FedAvg and $1 < \overline{K} \leq M/N$ (where $M/N = 20$ for the FMNIST case) implies cyclic client participation. A higher $\overline{K}$ represents the server visiting more client groups within a single cycle. We show that for high data heterogeneity, for all different client local procedures, a higher $\overline{K}$ achieves better test accuracy by approximately 5-10% improvement.



(a) GD (b) SGD (c) SSGD

Figure 2. Test accuracy for FMNIST for high data heterogeneity ($\alpha = 0.5$). CyCP ($\overline{K} > 1$) shows a higher test accuracy performance of 5-10% improvement compared to $\overline{K} = 1$ (Standard FedAvg) for all different client local procedures.

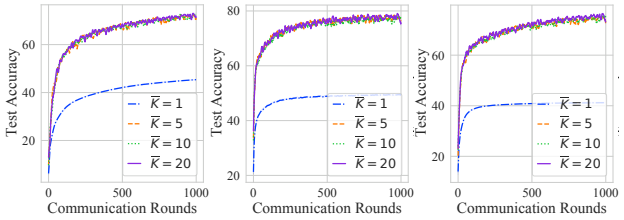


(a) GD (b) SGD (c) SSGD

Figure 3. Test accuracy for FMNIST for low data heterogeneity ($\alpha = 2.0$). Being consistent with the high data heterogeneity case in Fig. 2, CyCP ($\overline{K} > 1$) shows a higher test accuracy performance compared to $\overline{K} = 1$ (Standard FedAvg) for all different client local procedures with the improvement of 2-8%. However, the performance gap between CyCP and standard FedAvg is lower than when there is higher data heterogeneity.

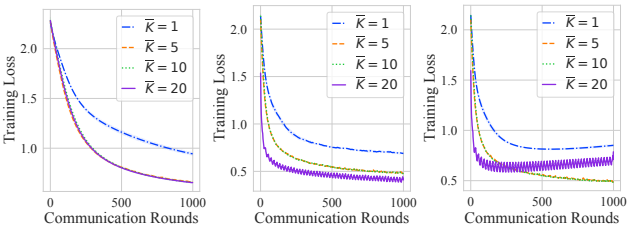
Although our theoretical results suggest that CyCP sees improvement in convergence for only $\overline{K} = M/N$, we observe improvement even for $M/N > \overline{K} > 1$. This can be due to our theoretical results being on PL-objectives while the landscape of DNN may not necessarily fall into this category (Qu et al., 2021). For lower data heterogeneity results shown in Fig. 3, the improvement for $\overline{K} > 1$ compared to standard FedAvg is approximately 2-8%. The improvement is less than that for the high data heterogeneity case which aligns with the theoretical results in Section 4 which shows that increasing $\overline{K}$ decreases the dominant term that is dependent on the intra-group data heterogeneity. In Fig. 4, we show that the performance gap between CyCP and standard FedAvg is even higher due to the data heterogeneity being even higher than the FMNIST case.

Difference Across Client Local Procedures. One distinct characteristic that can be observed in Fig. 2(b)-(c) and Fig. 4(b)-(c) which are the results for the local SGD and SSGD with high data heterogeneity is that for the highest $\overline{K}$ there are oscillations in the test accuracy curve. This is due to cyclic client participation where as we increase $\overline{K}$, the inter-group data heterogeneity also becomes higher causing oscillation as the server sequentially visits the group for training. This behavior has also been observed in previous



(a) GD (b) Local SGD (c) Shuffled SGD

Figure 4. Test accuracy for EMNIST with high data heterogeneity ($\alpha = 0.05$). CyCP shows a higher test accuracy performance to Standard FedAvg for all different client local procedures. The improvement gap is larger than the FMNIST case which is due to the EMNIST dataset partitioned with more data heterogeneity.



(a) GD (b) SGD (c) SSGD

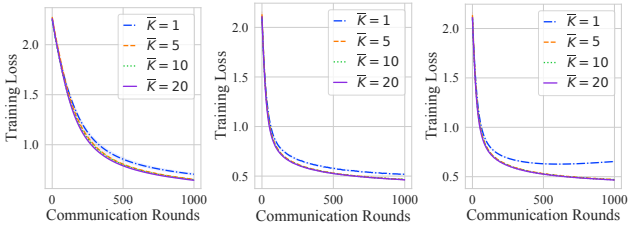
Figure 5. Training loss curves for the test accuracy results in Fig. 2 (FMNIST for high data heterogeneity, $\alpha = 0.5$) where higher $\overline{K} > 1$ leads to faster convergence, consistent with the test accuracy results where $\overline{K} > 1$ yields better performance.

work where the server trains sequentially in a cyclic manner from different groups in (Zhu et al., 2021). Accordingly, we show similar behavior of oscillation in the training loss curves shown in Fig. 5 and Fig. 6.

Training Losses for the Results in Fig. 2 & Fig. 3. We present the training loss for the test accuracy results shown in Fig. 2 and Fig. 3 in Fig. 5 and Fig. 6 respectively. We see that the implications are consistent with what we observed for test accuracy where a higher $\overline{K} > 1$ leads to faster convergence. The convergence improvement gap is larger when we have high data heterogenetiy as shown in Fig. 5 compared to the improvement gap for lower data heterogeneity shown in Fig. 6. Moreover, for the SGD and SSGD client local procedures for high data heterogeneity (Fig. 5(b)-(c)) for the highest $\overline{K} = 20$, we see the oscilliations that were also observed in the test accuracy curves. This is due to the cyclic participation of clients which the client groups have heterogeneous data. Such oscilliation is not observed for lower data heterogeneity and lower $\overline{K}$ values.

6. Concluding Remarks

Cyclic client participation is frequently observed in practical FL systems (Kairouz et al., 2021b; Zhu et al., 2021), but its effect on the convergence of FedAvg is not yet well-understood. In this paper, we formulate a new framework to



(a) GD ($\alpha = 2.0$) (b) SGD ($\alpha = 2.0$) (c) SSGD ($\alpha = 2.0$)

Figure 6. Training loss curves for the test accuracy results in Fig. 3 (FMNIST for low data heterogeneity, $\alpha = 2.0$) where higher $\overline{K} > 1$ also leads to faster convergence, consistent with the test accuracy results, but the convergence improvement gap is smaller compared to when we have higher data-heterogeneity.

analyze the convergence of FedAvg with cyclic client participation for PL-objectives. Our analysis covers different client local procedures such as GD, SGD, and shuffled SGD. The analysis allows us to understood how FedAvg convergence is affected by different characteristics of the system such local update procedures, data heterogeneity within and across groups of clients that cyclically become available, and the length of the cycle $\overline{K}$. We discover conditions in which cyclic client participation converges faster than standard FedAvg. We also provide comparisons across different client local procedures and algorithms that are special cases of our framework with cost analyses to achieve an ϵ error. Interesting future work includes extending the analysis to non-PL-objectives and general optimizers such as including momentum, and to cases where the client groups are not disjoint and not fixed throughout training.

Acknowledgements

This work was generously supported in part by NSF grants CCF 2045694, CNS-2112471, ONR N00014-23-1- 2149, and the Google Research Award.

References

Ahn, K., Yun, C., and Sra, S. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33:17526–17535, 2020.

Avdyukhin, D. and Kasiviswanathan, S. P. Federated learning under arbitrary communication patterns. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Balle, B., Kairouz, P., McMahan, B., Thakkar, O., and Guha Thakurta, A. Privacy amplification via random check-ins. *Advances in Neural Information Processing Systems*, 33:4623–4634, 2020.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations. In *Advances in Neural Information Processing Systems*, pp. 14695–14706, 2019.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. Towards Federated Learning at Scale: System Design. *SysML*, April 2019. URL https://www.sysml.cc/doc/2019/193.pdf.

Cho, Y. J., Wang, J., and Joshi, G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv:2010.01243*, abs/2010.01243, 2020. URL http://arxiv.org/abs/2010.01243.

Choquette-Choo, C. A., McMahan, H. B., Rush, K., and Thakurta, A. Multi-epoch matrix factorization mechanisms for private machine learning. *arXiv preprint arXiv:2211.06530*, 2022.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

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

Goetz, J., Malik, K., Bui, D., Moon, S., Liu, H., and Kumar, A. Active federated learning. *ArXiv*, 2019.

Gower, R. M., Sebbouh, O., and Loizou, N. SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Convergence rate of incremental gradient and incremental newton methods. *SIAM Journal on Optimization*, 29(4): 2542–2565, 2019.

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

Haddadpour, F. and Mahdavi, M. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, pp. 11080–11092, 2019.

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

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. In *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with NeurIPS 2019 (FL-NeurIPS'19)*, December 2019.

Huba, D., Nguyen, J., Malik, K., Zhu, R., Rabbat, M., Yousefpour, A., Wu, C.-J., Zhan, H., Ustinov, P., Srinivas, H., et al. Papaya: Practical, private, and scalable federated learning. *Proceedings of Machine Learning and Systems*, 4:814–832, 2022.

Jhunjhunwala, D., Sharma, P., Nagarkatti, A., and Joshi, G. Fedvarp: Tackling the variance due to partial client participation in federated learning. *arXiv*, 2022.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascon, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecny, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Ozgur, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramer, F., Vepakomma, P.,

Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021a.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. *arXiv preprint arXiv:2103.00039*, December 2021b.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. *CoRR*, abs/1608.04636, 2020. URL http://arxiv.org/abs/1608.04636.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of 37th International Conference on Machine Learning*, 2020.

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

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations (ICLR)*, July 2020b. URL https://arxiv.org/abs/1907.02189.

Li, X., Milzarek, A., and Qiu, J. Convergence of random reshuffling under the kurdyka-{\L} ojasiewicz inequality. *arXiv preprint arXiv:2110.04926*, 2021.

Malinovsky, G., Sailanbayev, A., and Richtárik, P. Random reshuffling with variance reduction: New analysis and better rates. *arXiv preprint arXiv:2104.09342*, 2021.

Malinovsky, G., Horváth, S., Burlachenko, K., and Richtárik, P. Federated learning with regularized client participation. *arXiv preprint arXiv:2302.03662*, 2023.

McMahan, B. and Thakurta, A. Federated learning with formal differential privacy guarantees. *Google AI Blog*, 2022.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. *International Conference on Artificial Intelligenece and Statistics (AISTATS)*, April 2017. URL https://arxiv.org/abs/1602.05629.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. *International Conference on Learning Representations*, 2018.

Mishchenko, K., Khaled, A., and Richtárik, P. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33: 17309–17320, 2020.

Nagaraj, D., Jain, P., and Netrapalli, P. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pp. 4703–4711. PMLR, 2019.

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and Van Dijk, M. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

Paulik, M., Seigel, M., Mason, H., Telaar, D., Kluivers, J., van Dalen, R., Lau, C. W., Carlson, L., Granqvist, F., Vandevelde, C., et al. Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv preprint arXiv:2102.08503*, 2021.

Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., and Zhou, Z. Federated learning's blessing: Fedavg has linear speedup. *https://arxiv.org/abs/2007.05690*, 2020.

Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., and Zhou, Z. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *https://arxiv.org/abs/2003.00307*, 2021.

Rajput, S., Gupta, A., and Papailiopoulos, D. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, 2020.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Ruan, Y., Zhang, X., Liang, S.-C., and Joe-Wong, C. Towards flexible device participation in federated learning for non-iid data. *ArXiv*, 2020.

Safran, I. and Shamir, O. How good is sgd with random shuffling? In *Conference on Learning Theory*, pp. 3250–3284. PMLR, 2020.

Safran, I. and Shamir, O. Random shuffling beats sgd only after many epochs on ill-conditioned problems. *Advances in Neural Information Processing Systems*, 34:15151–15161, 2021.

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. Federated optimization for heterogeneous networks. In *Proceedings of the 3rd MLSys Conference*, January 2020.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *preprint*, May 2020a. URL https://arxiv.org/abs/2007.07481.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. SlowMo: Improving communication-efficient distributed SGD with slow momentum. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=SkxJ8REYPH.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

Wang, S. and Ji, M. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, 2022.

Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. Is local SGD better than minibatch SGD? In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

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

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *https://arxiv.org/abs/1708.07747*, aug 2017.

Xu, Z., Zhang, Y., Andrew, G., Choquette, C., Kairouz, P., McMahan, B., Rosenstock, J., and Zhang, Y. Federated learning of gboard language models with differential privacy, 2023.

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

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, jun 2019.

Yun, C., Rajput, S., and Sra, S. Minibatch vs local sgd with shuffling: Tight convergence bounds and beyond. *International Conference on Learning Representations (ICLR)*, 2022.

Zhu, C., Xu, Z., Chen, M., Konečnỳ, J., Hard, A., and Goldstein, T. Diurnal or nocturnal? federated learning of multi-branch networks from periodically shifting distributions. In *International Conference on Learning Representations*, 2021.

Zhu, F., Zhang, J., and Wang, X. Communication-efficient local sgd with age-based worker selection. *arXiv preprint arXiv:2210.17073*, December 2022.

A. Proofs for the CyCP+Local GD Case (Case (i) in Algorithm 1)

The proof of Theorem 1 is presented in this section. For simplicity, the proof is presented as follows: first, in appendix A.1, the model update steps are shown to be noisy gradient descent steps. Next, in appendix A.2, we present some intermediate results, which shall be used in the analysis, followed by the proof of Theorem 1 in appendix A.3. Finally, in appendix A.4, we present the proofs of the intermediate results.

We define the σ-algebra generated by the randomness in the algorithm till cycle epoch k as follows: $\mathcal{F}_k \triangleq \sigma\left\{\{\mathbf{w}^{(1,i)}\}_{i=1}^{\overline{K}}, \{\mathbf{w}^{(2,i)}\}_{i=1}^{\overline{K}}, \ldots, \{\mathbf{w}^{(k-1,i)}\}_{i=1}^{\overline{K}}\right\}$. We use $\mathbb{E}_k[\cdot]$ as the shorthand for the expectation $\mathbb{E}[\cdot|\mathcal{F}_k]$.

A.1. Global Model Updates as Noisy Gradient Descent Step

With CyCP, with each client doing a single GD update, (case (i) in Algorithm 1), the update rule for the global model ($k \in [K], i \in [\overline{K}]$) is given as

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)}) \tag{12}$$

Recall that $\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,\overline{K})}$. Therefore, we can unroll (12) to get the following result.

Lemma 2.

$$\mathbb{E}[\mathbf{w}^{(k+1,0)}|\mathcal{F}_k] - \mathbf{w}^{(k,0)} = -\eta\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \eta^2\mathbb{E}[\overline{\mathbf{r}}^{(k,0)}|\mathcal{F}_k], \tag{13}$$

where $\overline{\mathbf{q}}^{(k,i)} := \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)})$ *and* $\overline{\mathbf{S}}^{(k,i)} := \sum_{m \in \mathcal{S}^{(k,i)}} \mathbf{H}_m^{(k,i)}$ *we have* $\overline{\mathbf{r}}^{(k,0)} \triangleq \sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d - \eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}$.

Proof. Note that (12) is a rescaled version of the update rule in the main paper in the sense that the update is scaled up by N but instead we downscale the step-size by setting the learning rate as $\eta = \log(MK^2)/\mu MK$. We can reformulate the gradient of the local objective as:

$$\begin{aligned}
\nabla F_m(\mathbf{w}^{(k,i-1)}) &= \nabla F_m(\mathbf{w}^{(k,0)}) + \nabla F_m(\mathbf{w}^{(k,i-1)}) - \nabla F_m(\mathbf{w}^{(k,0)}) \\
\Rightarrow \nabla F_m(\mathbf{w}^{(k,i-1)}) - \nabla F_m(\mathbf{w}^{(k,0)}) &= \underbrace{\int_0^1 \nabla^2 F_m(\mathbf{w}^{(k,0)} + t(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}))\, dt}_{:=\mathbf{H}_m^{(k,i)}} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})
\end{aligned} \tag{14}$$

Note that $\mathbf{H}_m^{(k,i)}$ exists due to our assumption that the local objectives $F_m(\cdot),\ m \in [M]$ are differentiable and L-smooth (see Appendix D.2 in (Yun et al., 2022)), and we can thus show that $\|\mathbf{H}_m^{(k,i)}\| \leq L,\ \forall m, k, i$. Using the expression in (14) we get

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m \in \mathcal{S}^{(k,i)}} \left(\nabla F_m(\mathbf{w}^{(k,0)}) + \mathbf{H}_m^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})\right). \tag{15}$$

Defining $\overline{\mathbf{q}}^{(k,i)} := \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)})$ and $\overline{\mathbf{S}}^{(k,i)} := \sum_{m \in \mathcal{S}^{(k,i)}} \mathbf{H}_m^{(k,i)}$ we have

$$\begin{aligned}
\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} &= -\eta\overline{\mathbf{q}}^{(k,i)} - \eta\overline{\mathbf{S}}^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}) \\
\Rightarrow \mathbf{w}^{(k,i)} - \mathbf{w}^{(k,0)} &= \left(\mathbf{I} - \eta\overline{\mathbf{S}}^{(k,i)}\right)(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}) - \eta\overline{\mathbf{q}}^{(k,i)}.
\end{aligned} \tag{16}$$

Unrolling (16), we get

$$\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} = -\eta\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)} + \eta^2 \underbrace{\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d - \eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}}_{:=\overline{\mathbf{r}}^{(k,0)}} \tag{17}$$

Conditioning on $\mathcal{F}_k$, the only randomness in (17) is owing to the random client sets $\{\mathcal{S}^{(k,i)}\}_i$. Therefore,

$$\mathbb{E}[\mathbf{w}^{(k+1,0)}|\mathcal{F}_k] - \mathbf{w}^{(k,0)} = -\eta \sum_{i=1}^{\overline{K}} \mathbb{E}_k \left[\sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)}) \right] + \eta^2 \mathbb{E}_k \left[\overline{\mathbf{r}}^{(k,0)} \right],$$

Computing the expectation finishes the proof. □

Interpreting $\eta \overline{K} N$ as the effective learning rate, the update in (13) is a noisy gradient descent step. The bulk of the proof is concerned with bounding the noise term $\overline{\mathbf{r}}^{(k,0)}$.

A.2. Intermediate Results

Lemma 3. *If the client functions F_m satisfy Assumption 3, then for arbitrary $i \in [\overline{K}]$ and any $\mathbf{w}$*

$$\left\| \sum_{j=1}^{i} \frac{1}{M/\overline{K}} \sum_{m \in \sigma(j)} \nabla F_m(\mathbf{w}) \right\| \leq i\alpha + i\|\nabla F(\mathbf{w})\|.$$

Lemma 4 (Bounds on the Error Terms)**.** *Under Assumption 1, 3, the error in* (13) *can be bounded as*

$$\left\| \mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}] \right\| \leq \frac{3N^2 L \overline{K}(\overline{K}-1)}{5} \left(\alpha + \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \right), \tag{18}$$

$$\mathbb{E}_k \left[\left\| \overline{\mathbf{r}}^{(k,0)} \right\|^2 \right] \leq \frac{(\overline{K}-1)^2 \overline{K}^2 N^3 L^2}{2} \left[\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + 2N \left(\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \alpha^2 \right) \right]. \tag{19}$$

A.3. Proof of Theorem 1

For ease of reference, we restate Theorem 1 here.

Theorem (Convergence of the Global Model with CyCP+Local GD (Case (i) in Algorithm 1))**.** *Suppose the local client functions $\{F_m\}$ satisfy Assumption 1 and Assumption 3, while the global loss function satisfies Assumption 2. If the learning rate satisfies $\eta \leq \frac{1}{7LN\overline{K}}$, then the iterates generated by Algorithm 1 with one local GD step at the clients satisfies*

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \left(1 - \eta \overline{K} N \mu\right)^K \left(F(\mathbf{w}^{(0,0)}) - F^*\right) + \mathcal{O}\left(\frac{\eta^2 L^2 N^2 (\overline{K}-1)^2 \alpha^2}{\mu} \right) + \mathcal{O}\left(\kappa \eta \overline{K} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 \right),$$

where $\kappa = L/\mu$ is the condition number. Choosing step-size $\eta = \log(MK^2)/\mu N \overline{K} K$ with $K \geq 7\kappa \log(MK^2)$, the convergence error is bounded as:

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2 (\overline{K}-1)^2 \alpha^2}{\mu \overline{K}^2 K^2} \right) + \tilde{\mathcal{O}}\left(\frac{\kappa \gamma^2}{\mu N K} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \right)$$

where $\tilde{\mathcal{O}}(\cdot)$ subsumes logarithmic terms and numerical constants.

Corollary 1. *Denoting the total number of communication rounds in algorithm 1 as $T = \overline{K} K$, in terms of T the bound above becomes*

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2 (F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2 (\overline{K}-1)^2 \alpha^2}{\mu T^2} \right) + \tilde{\mathcal{O}}\left(\frac{\overline{K} \kappa \gamma^2}{\mu N T} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \right).$$

where since $K \geq 7\kappa \log(MK^2)$, we have $T \geq 7\kappa \overline{K} \log(M \overline{K}^2 T^2)$ and one cannot increase $\overline{K}$ without increasing T accordingly due to its lower bound depending on $\overline{K}$.

Proof. Using the L-smoothness property (Assumption 1) of the global objective $F(\mathbf{w})$ we have

$$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq \langle \nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} \rangle + \frac{L}{2} \mathbb{E}_k \left[\left\| \mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} \right\|^2 \right] \tag{20}$$

where the expectation here is over the selected client sets $\mathcal{S}^{(k,i)}$, for all $i \in [\overline{K}]$. First, we bound the inner product term in (20). Using Lemma 2, Lemma 4, we have

$$\begin{aligned}
&\langle \nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} \rangle \\
&= \langle \nabla F(\mathbf{w}^{(k,0)}), -\eta \overline{K} N \nabla F(\mathbf{w}^{(k,0)}) + \eta^2 \mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}] \rangle && \text{(Using Lemma 2)} \\
&\leq -\eta \overline{K} N \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \eta^2 \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \frac{3N^2 L \overline{K}(\overline{K}-1)}{5} \left(\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \alpha \right) && \text{(Using Lemma 4)} \\
&= -\eta \overline{K} N \left(1 - \frac{3\eta(\overline{K}-1)NL}{5} \right) \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{3\eta^2 N^2 L \overline{K}(\overline{K}-1)\alpha}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| && (21)
\end{aligned}$$

We can bound the last term in (21) as

$$\begin{aligned}
\frac{3\eta^2 N^2 L \overline{K}(\overline{K}-1)\alpha}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| &= \left(\frac{\eta^{1/2} N^{1/2} \overline{K}^{1/2}}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \right) \left(3\eta^{3/2} N^{3/2} L \overline{K}^{1/2} (\overline{K}-1)\alpha \right) \\
&\leq \frac{\eta N \overline{K}}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{9\eta^3 \alpha^2 L^2 N^3 \overline{K}(\overline{K}-1)^2}{2} && (22)
\end{aligned}$$

Plugging (22) back into (21) we have

$$\langle \nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} \rangle \leq -\eta \overline{K} N \left(\frac{49}{50} - \frac{3\eta(\overline{K}-1)NL}{5} \right) \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{9\eta^3 \alpha^2 L^2 N^3 \overline{K}(\overline{K}-1)^2}{2}. \tag{23}$$

Now we aim to bound the last term in the RHS of (20). We have

$$\frac{L}{2} \mathbb{E}_k \left[\left\| \mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} \right\|^2 \right] \overset{(17)}{=} \frac{L}{2} \mathbb{E}_k \left[\left\| -\eta \sum_{i=1}^{\overline{K}} \overline{\mathbf{q}}^{(k,i)} + \eta^2 \overline{\mathbf{r}}^{(k,0)} \right\|^2 \right] \leq L\eta^2 \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \overline{\mathbf{q}}^{(k,i)} \right\|^2 \right] + L\eta^4 \mathbb{E}_k \left[\left\| \overline{\mathbf{r}}^{(k,0)} \right\|^2 \right] \tag{24}$$

The last term is already bounded in Lemma 4, (19). Now, we bound the first term,

$$\begin{aligned}
&\mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \overline{\mathbf{q}}^{(k,i)} \right\|^2 \right] = \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)}) \right\|^2 \right] = N^2 \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)}) \right\|^2 \right] \\
&= N^2 \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \left\{ \frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)}) - \frac{1}{M/\overline{K}} \sum_{m \in \sigma(i)} \nabla F_m(\mathbf{w}^{(k,0)}) + \frac{1}{M/\overline{K}} \sum_{m \in \sigma(i)} \nabla F_m(\mathbf{w}^{(k,0)}) \right\} \right\|^2 \right] \\
&= N^2 \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}} \left\{ \frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)}) - \frac{1}{M/\overline{K}} \sum_{m \in \sigma(i)} \nabla F_m(\mathbf{w}^{(k,0)}) \right\} \right\|^2 \right] + N^2 \left\| \frac{1}{M/\overline{K}} \sum_{i=1}^{\overline{K}} \sum_{m \in \sigma(i)} \nabla F_m(\mathbf{w}^{(k,0)}) \right\|^2 \\
&\leq N \overline{K}^2 \left[\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + N \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 \right]. && (25)
\end{aligned}$$

following steps analogous to the proof of (36). Finally, we can plug in (23), (24) and (25) into (20) to get

$$\begin{aligned}
&\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \\
&\leq -\eta \overline{K} N \left(\frac{49}{50} - \frac{3\eta(\overline{K}-1)NL}{5} - L\eta N \overline{K} - \eta^3 \overline{K}(\overline{K}-1)^2 N^3 L^3 \right) \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{9\eta^3 \alpha^2 L^2 N^3 \overline{K}(\overline{K}-1)^2}{2} \\
&+ \eta^4 \overline{K}^3 (\overline{K}-1) N^4 L^3 \alpha^2 + \frac{1}{2} \eta^4 \overline{K}^2 (\overline{K}-1)^2 N^3 L^3 \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + L\eta^2 N \overline{K}^2 \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2
\end{aligned}$$

$$\leq -\frac{\eta\overline{K}N}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10}\alpha^2 + \frac{11L\eta^2N\overline{K}^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$

$$\leq (1-\eta\overline{K}N\mu)(\mathbb{E}[F(\mathbf{w}^{(k,0)})]-F^*) + \frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10}\alpha^2 + \frac{11L\eta^2N\overline{K}^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2, \tag{26}$$

where the last inequality follows from Assumption 2. Unrolling (26), and using $\eta = \frac{\log(MK^2)}{\mu\overline{K}KN}$, we get

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \\
&\leq \left(1-\frac{\log(MK^2)}{K}\right)^K (F(\mathbf{w}^{(0,0)})-F^*) + \frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10\eta\overline{K}N\mu}\alpha^2 + \frac{11L\eta^2N\overline{K}^2}{5\eta\overline{K}N\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \\
&\leq \frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2} + \frac{47\eta^2L^2N^2(\overline{K}-1)^2}{10\mu}\alpha^2 + \frac{11L\eta\overline{K}}{5\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \\
&= \frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2} + \frac{47\log^2(MK^2)\kappa^2(\overline{K}-1)^2}{10\mu\overline{K}^2K^2}\alpha^2 + \frac{11L\log(MK^2)}{5\mu^2NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \\
&= \frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right) + \tilde{\mathcal{O}}\left(\frac{\kappa}{\mu NK}\left(\frac{M-N\overline{K}}{M-\overline{K}}\right)\gamma^2\right).
\end{aligned}$$

Also, since we have total $T = \overline{K}K$ communication rounds, we can also express the bound in terms of T as follows.

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)})-F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa}{\mu NT}\left(\frac{M-N\overline{K}}{M-\overline{K}}\right)\gamma^2\right).$$

□

Proof for Corollary 1 We reiterate (3), the total cost C_{GD} to achieve an ϵ error:

$$C_{\text{GD}}(\epsilon) = \frac{c_{\text{GD}}\overline{K}\gamma^2}{\epsilon N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right).$$

To get $C_{\text{GD}|\overline{K}>1}(\epsilon) < C_{\text{GD}|\overline{K}=1}(\epsilon)$ we have that the inequality

$$\overline{K}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) - \left(\frac{M-N}{M-1}\right) < 0$$

should be true for $\overline{K} > 1$. Rearranging the terms, we get

$$N(M-1)\overline{K}^2 + (N-M^2)\overline{K} + M^2 - MN > 0 \tag{27}$$

Since $(M(M-2N)+N)^2 = (N-M^2)^2 - 4N(M^2-MN)(M-1) \geq 0$ we have that for $\overline{K} > M(M-N)/N(M-1)$ we have $\overline{K}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) - \left(\frac{M-N}{M-1}\right) < 0$ Rearranging the terms slightly, we get

$$M < N + N\overline{K}\left(1-\frac{1}{M}\right).$$

Recall that $\overline{K} \in [\frac{M}{N}]$. For $\overline{K} = M/N$, we get

$$N + N\frac{M}{N}\left(1-\frac{1}{M}\right) = M-1+N > M$$

for any $N > 1$. However, if $\overline{K} = \frac{M}{N} - 1$, we get

$$N + N\left(\frac{M}{N}-1\right)\left(1-\frac{1}{M}\right) = M-1+N-N+\frac{N}{M} = M-1+\frac{N}{M} < M$$

for any $N > 1$. Consequently, the only case when $\overline{K} > 1$ gives benefit over $\overline{K} = 1$ is when $\overline{K} = M/N$, meaning full client participation in every client group.

A.4. Proofs on Intermediate Lemmas

Proof of Lemma 3.

$$\left\|\sum_{j=1}^{i}\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w})\right\| \leq \sum_{j=1}^{i}\left\|\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w})-\nabla F(\mathbf{w})+\nabla F(\mathbf{w})\right\|$$
$$\leq \sum_{j=1}^{i}\left\|\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w})-\nabla F(\mathbf{w})\right\|+i\|\nabla F(\mathbf{w})\| \leq i\alpha+i\|\nabla F(\mathbf{w})\|. \qquad \text{(using Assumption 3)}$$

□

Proof of Lemma 4. First, we bound $\left\|\mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}]\right\|$.

$$\left\|\mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}]\right\| = \left\|\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right]\right\|$$
$$\leq \sum_{i=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right]\right\| \tag{28}$$
$$\leq \sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\underbrace{\left\|\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|}_{A_1}\underbrace{\left\|\overline{\mathbf{S}}^{(k,i+1)}\right\|}_{A_2}\underbrace{\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|}_{A3}\right] \tag{29}$$

Next, we bound A_1, A_2, A_3 separately as follows.

$$A_1 = \left\|\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\| \leq \prod_{j=i+2}^{\overline{K}}\left\|(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|$$
$$= \prod_{j=i+2}^{\overline{K}}\left\|I_d-\eta\sum_{m\in\mathcal{S}^{(k,j)}}\mathbf{H}_m^{(k,j)}\right\| \leq (1+\eta NL)^{\overline{K}} \leq e^{1/7} \leq 6/5, \tag{30}$$

where in (30) we use Assumption 1, and $\eta \leq \frac{1}{7LN\overline{K}}$. Next, we bound A_2.

$$A_2 = \left\|\overline{\mathbf{S}}^{(k,i+1)}\right\| = \left\|\sum_{m\in\mathcal{S}^{(k,i+1)}}\mathbf{H}_m^{(k,i+1)}\right\| \leq NL \tag{31}$$

Next, we bound A_3 as follows:

$$A_3 = \left\|\sum_{j'=1}^{i}\mathbb{E}_k\left[\overline{\mathbf{q}}^{(k,j')}\right]\right\| = \left\|\sum_{j'=1}^{i}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})\right]\right\| = N\left\|\sum_{j'=1}^{i}\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|$$
$$\leq iN\alpha+iN\|\nabla F(\mathbf{w})\|. \tag{32}$$

where the last inequality follows from Lemma 3. Substituting the bounds from (30)-(32) in (29), we get the bound in (18). Next, we derive the bound in (19).

$$\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right] = \mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]$$

$$\leq (\overline{K}-1)\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]$$

$$\leq (\overline{K}-1)\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|^2\left\|\overline{\mathbf{S}}^{(k,i+1)}\right\|^2\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]. \tag{33}$$

Observe that

$$\left\|(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|\leq 1+\eta\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{H}_m^{(k,i)}\right\|\leq 1+\eta\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\mathbf{H}_m^{(k,i)}\right\|\leq 1+\eta NL,\ \forall j\in[\overline{K}]. \tag{34}$$

Using (30), (31) and (34) in (33), we get

$$\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right]\leq\frac{36(\overline{K}-1)N^2L^2}{25}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right] \tag{35}$$

Lastly, we bound the expected value in (35):

$$\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left\|\sum_{j'=1}^{i}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})+\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2$$

$$=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2\right]$$
$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$

$$\leq N^2\sum_{i=1}^{\overline{K}-1}i\sum_{j'=1}^{i}\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]$$
$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$

$$=N^2\sum_{i=1}^{\overline{K}-1}i\sum_{j'=1}^{i}\frac{1}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\frac{1}{M/\overline{K}}\sum_{m'\in\sigma(j')}\left\|\nabla F_{m'}(\mathbf{w})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w})\right\|^2$$
$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$

(Using without replacement sampling)

$$\leq N^2\sum_{i=1}^{\overline{K}-1}i\sum_{j'=1}^{i}\frac{1}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \qquad \text{(Assumption 3)}$$

$$\leq \sum_{i=1}^{\overline{K}-1} i^2 N \left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + N^2 \sum_{i=1}^{\overline{K}-1} 2i^2 \left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \alpha^2\right) \quad \text{(using Lemma 3)}$$

$$\leq \frac{1}{3}(\overline{K}-1)\overline{K}^2 N \left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \alpha^2\right)\right]. \tag{36}$$

Plugging (36) into (35) we get

$$\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right] \leq \frac{12(\overline{K}-1)^2\overline{K}^2 N^3 L^2}{25}\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \alpha^2\right)\right].$$

which concludes the proof. □

B. Proofs for the CyCP+Local SGD Case (Case (ii) in Algorithm 1)

The proof of Theorem 1 is presented in this section. For simplicity, the proof is presented as follows: first, in appendix B.1, the model update steps are shown to be noisy gradient descent steps. Next, in appendix B.2, we present some intermediate results, which shall be used in the analysis, followed by the proof of Theorem 2 in appendix B.3. Finally, in appendix B.5, we present the proofs of the intermediate results.

Similar to what we have did in appendix A.3, we define the σ-algebra generated by the randomness in the algorithm till cycle epoch k as follows: $\mathcal{F}_k \triangleq \sigma\left\{\{\mathbf{w}^{(1,i)}\}_{i=1}^{\overline{K}}, \{\mathbf{w}^{(2,i)}\}_{i=1}^{\overline{K}}, \ldots, \{\mathbf{w}^{(k-1,i)}\}_{i=1}^{\overline{K}}\right\}$. We use $\mathbb{E}_k[\cdot]$ as the shorthand for the expectation $\mathbb{E}[\cdot|\mathcal{F}_k]$.

B.1. Global Model Updates as Noisy Gradient Descent Step

Recall that the global model update rule is

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) \tag{37}$$

where $\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) = \frac{1}{b}\sum_{\xi\in\xi_m^{(k,i-1,l)}} \nabla f(\mathbf{w}_m^{(k,i-1,l)}, \xi)$ is the stochastic gradient computed using a mini-batch $\xi_m^{(k,i-1,l)}$ of size b that is randomly sampled from client m's local dataset $\mathcal{B}_m$. Recall that k is a *semi-epoch* index which denotes each communication round when we have traversed the $\overline{K}$ groups of clients by sampling N clients from each group $\sigma(i),\ i\in[\overline{K}]$ uniformly at random without replacement. The index $i\in[\overline{K}]$ denotes each inner communication round. Recall that $\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,\overline{K})}$ for all $k\in[K]$. With (37) we get

$$\begin{aligned}
\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} &= -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \Big(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)}) + \nabla F_m(\mathbf{w}^{(k,i-1)})\Big) \\
&= -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \underbrace{\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)}))}_{:=\mathbf{d}_m^{(k,i)}} - \eta\tau \sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)}) \\
&= -\eta \underbrace{\sum_{m\in\mathcal{S}^{(k,i)}} \mathbf{d}_m^{(k,i)}}_{:=\overline{\mathbf{d}}^{(k,i)}} - \eta\tau \sum_{m\in\mathcal{S}^{(k,i)}} \Big(\nabla F_m(\mathbf{w}^{(k,0)}) + \mathbf{H}_m^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}\Big) \qquad \text{(from (14))} \\
&= -\widetilde{\eta}\underbrace{(\overline{\mathbf{d}}^{(k,i)}/\tau + \overline{\mathbf{q}}^{(k,i)})}_{:=\widetilde{\mathbf{q}}^{(k,i)}} - \widetilde{\eta}\underbrace{\sum_{m\in\mathcal{S}^{(k,i)}} \mathbf{H}_m^{(k,i)}}_{:=\overline{\mathbf{S}}^{(k,i)}}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})
\end{aligned} \tag{38}$$

where $\widetilde{\eta} = \tau\eta$ and $\overline{\mathbf{q}}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)})$. Unrolling (38), we get

$$\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} = -\widetilde{\eta}\sum_{i=1}^{\overline{K}} \widetilde{\mathbf{q}}^{(k,i)} + \widetilde{\eta}^2 \underbrace{\sum_{i=1}^{\overline{K}-1} \left(\prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right) \overline{\mathbf{S}}^{(k,i+1)} \sum_{j'=1}^{i} \widetilde{\mathbf{q}}^{(k,j')}}_{:=\widetilde{\mathbf{r}}^{(k,0)}} \tag{39}$$

Conditioning on $\mathcal{F}_k$, we get

$$\begin{aligned}
\mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} &= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}} \mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right] - \widetilde{\eta}\mathbb{E}_k\left[\sum_{i=1}^{\overline{K}} \sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)})\right] + \widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}] \\
&= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}} \mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right] - \widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]
\end{aligned}$$

$$= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right] - \widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]. \tag{40}$$

B.2. Intermediate Results

Lemma 5. *First to bound the first term in (42) we bound the following:*

$$\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\| \leq \frac{3\eta LN\overline{K}(\tau-1)\tau}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{14\eta LN\overline{K}(\tau-1)\tau\nu}{25} + \frac{\eta LN(\overline{K}-1)(\tau-1)\tau\alpha}{36}.$$

Lemma 6.

$$\widetilde{\eta}^2\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\| \leq \frac{17\widetilde{\eta}N\overline{K}}{250}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{84\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)\nu}{125} + \frac{16\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{25}.$$

Lemma 7.

$$\begin{aligned}
&\sum_{r=1}^{\overline{K}}\mathbb{E}\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right] \\
&\leq \frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau} + \frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100} \\
&+ \frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}.
\end{aligned}$$

Lemma 8.

$$\begin{aligned}
\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right] &\leq \frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau} + \frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50} \\
&+ \frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100}.
\end{aligned}$$

Lemma 9.

$$\begin{aligned}
\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right] &\leq \frac{9\widetilde{\eta}N\overline{K}}{500}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{3\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{200\tau} + \frac{173\widetilde{\eta}^4\overline{K}^4N^3L^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \\
&+ \frac{3\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50} + \frac{59\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2}{10}.
\end{aligned}$$

B.3. Proof of Theorem 2

Again, with the L-smoothness property of the global objective we have that

$$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq \langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle + \frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right] \tag{41}$$

where the expectation here is over the selected client sets $\mathcal{S}^{(k,i)},\ i\in[\overline{K}]$ and stochastic gradients. First we bound the inner product term in the RHS (41). Using (40),

$$\begin{aligned}
&\langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle \\
&= \langle\nabla F(\mathbf{w}^{(k,0)}), -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right] - \widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\rangle
\end{aligned}$$

$$\leq \frac{\widetilde{\eta}}{\tau}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\| - \widetilde{\eta}\overline{K}N\left\|\nabla F(\mathbf{w}^{(k,0)}\right\|^2 + \widetilde{\eta}^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\|. \tag{42}$$

$\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\|$ and $\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\|$ are already bounded in Lemma 5 and Lemma 6 respectively. Plugging these bounds in (42) we have

$$\begin{aligned}
&\langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle \\
&\leq \frac{\widetilde{\eta}}{\tau}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left(\frac{3\eta LN\overline{K}(\tau-1)\tau}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{14\eta LN\overline{K}(\tau-1)\tau\nu}{25} + \frac{\eta LN(\overline{K}-1)(\tau-1)\tau\alpha}{36}\right) \\
&\quad - \widetilde{\eta}\overline{K}N\left\|\nabla F(\mathbf{w}^{(k,0)}\right\|^2 + \left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left(\frac{17\widetilde{\eta}N\overline{K}}{250}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{84\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)\nu}{125}\right. \\
&\qquad\qquad \left. + \frac{16\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{25}\right) \\
&\leq -\widetilde{\eta}\overline{K}N\left(1-\frac{3}{50}-\frac{17}{250}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)(14\widetilde{\eta}^{1/2}\eta LN^{1/2}\overline{K}^{1/2}(\tau-1)\nu) \\
&\quad + \left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{40}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)(28\widetilde{\eta}^{3/2}\eta L^2N^{3/2}\overline{K}^{3/2}(\tau-1)\nu) \\
&\quad + \left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{36}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{\widetilde{\eta}^{1/2}\eta LN^{1/2}(\overline{K}-1)(\tau-1)\alpha}{\overline{K}^{1/2}}\right) \\
&\quad + \left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(16\widetilde{\eta}^{3/2}LN^{3/2}\overline{K}^{1/2}(\overline{K}-1)\alpha\right) \\
&\leq -\widetilde{\eta}\overline{K}N\left(1-\frac{3}{50}-\frac{17}{250}-\frac{1}{2\times 40^2}-\frac{1}{2\times 36^2}-\frac{1}{2\times 25^2}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + 98\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)^2\nu^2 \\
&\quad + 392\widetilde{\eta}^3\eta^2L^4N^3\overline{K}^3(\tau-1)^2\nu^2 + \frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}} + 128\widetilde{\eta}^3L^2N^3\overline{K}(\overline{K}-1)^2\alpha^2 \\
&\leq -\frac{107\widetilde{\eta}\overline{K}N}{125}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + 102\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)^2\nu^2 \\
&\quad + \frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}} + 128\widetilde{\eta}^3L^2N^3\overline{K}(\overline{K}-1)^2\alpha^2.
\end{aligned} \tag{43}$$

Now we bound the second term in the RHS of (41) as follows

$$\begin{aligned}
\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right] &= \frac{L}{2}\mathbb{E}_k\left[\left\|-\widetilde{\eta}\sum_{i=1}^{\overline{K}}\widetilde{\mathbf{q}}^{(k,i)} + \widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right] \\
&= \frac{L}{2}\mathbb{E}_k\left[\left\|-\widetilde{\eta}\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}/\tau - \widetilde{\eta}\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)} + \widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right] \qquad \text{(Using (39))} \\
&\leq \frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right] + \frac{3L\widetilde{\eta}^2}{2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right] + \frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right].
\end{aligned} \tag{44}$$

Substituting the bounds from (25), Lemma 8 and Lemma 9 into (44) we get

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right] \leq \frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau} + \frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50} \\
&+ \frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100} + \frac{3L\widetilde{\eta}^2N\gamma^2\overline{K}^2}{2}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \\
&+ \frac{3\widetilde{\eta}N\overline{K}}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{9\widetilde{\eta}N\overline{K}}{500}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{3\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{200\tau} + \frac{173\widetilde{\eta}^4\overline{K}^4N^3L^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \\
&+ \frac{3\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50} + \frac{59\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2}{10} \\
&\leq \widetilde{\eta}N\overline{K}\left(\frac{1}{200}+\frac{3}{20}+\frac{9}{500}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{\tau}\left(\frac{81}{50}+\frac{3}{200}\right) \\
&+ (\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2)\left(\frac{209}{50}+\frac{3}{50}\right) + L\widetilde{\eta}^2\overline{K}^2N\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(\frac{174}{2000}+\frac{3}{2}\right) \\
&+ L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2\left(\frac{9}{100}+\frac{59}{10}\right) \\
&\leq \frac{173\widetilde{\eta}N\overline{K}}{1000}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau} + \frac{41\widetilde{\eta}\eta^2L^2\overline{K}N\tau(\tau-1)\nu^2}{250} + \frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \qquad (45) \\
&+ 6L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2
\end{aligned}
$$

Using the bound derived in (45) and (43) for the upper bound in (41) we have

$$
\begin{aligned}
\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq &-\frac{\widetilde{\eta}\overline{K}N}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + 103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2 + \frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}} \\
&+ 129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2 + \frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau} + \frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \qquad (46)
\end{aligned}
$$

With Assumption 2 we have

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(k+1,0)})] - F^* \leq &(1-\widetilde{\eta}\overline{K}N\mu)(\mathbb{E}[F(\mathbf{w}^{(k,0)})] - F^*) + 103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2 + \frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}} \\
&+ 129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2 + \frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau} + \frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right), \qquad (47)
\end{aligned}
$$

and unrolling (47) we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \left(1-\frac{\log(MK^2)}{K}\right)^K(F(\mathbf{w}^{(0,0)}) - F^*) + \frac{103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2}{\widetilde{\eta}\overline{K}N\mu} + \frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)(\tau-1)^2\alpha^2}{2\widetilde{\eta}\overline{K}N\mu} \\
&+ \frac{129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2}{\widetilde{\eta}\overline{K}N\mu} + \frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau\widetilde{\eta}\overline{K}N\mu} + \frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5\widetilde{\eta}\overline{K}N\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \\
&\leq \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \frac{103\eta^2L^2(\tau-1)\tau\nu^2}{\mu} + \frac{\eta^2L^2(\overline{K}-1)(\tau-1)^2\alpha^2}{2\overline{K}\mu} + \frac{129\widetilde{\eta}^2L^2N^2\overline{K}(\overline{K}-1)\alpha^2}{\mu} \\
&+ \frac{41L\widetilde{\eta}\overline{K}\sigma^2}{25\tau\mu} + \frac{8L\widetilde{\eta}\overline{K}\gamma^2}{5\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) \\
&= \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \frac{103\log^2(MK^2)\kappa^2(\tau-1)\nu^2}{\mu\tau N^2\overline{K}^2K^2} + \frac{\log^2(MK^2)\kappa^2(\overline{K}-1)(\tau-1)\alpha^2}{2\mu\tau N^2\overline{K}^3K^2} \\
&+ \frac{196\log^2(MK^2)\kappa^2(\overline{K}-1)\alpha^2}{\mu\overline{K}K^2} + \frac{41\log(MK^2)\kappa\sigma^2}{25\mu\tau NK} + \frac{8\log(MK^2)\kappa\gamma^2}{5\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)
\end{aligned}
$$

$$= \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2\overline{K}^2K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)(\overline{K}-1)\alpha^2}{\mu\tau N^2\overline{K}^3K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)\alpha^2}{\mu\overline{K}K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{\mu\tau NK}\right)$$
$$+ \widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) \tag{48}$$

B.4. Proof for Corollary 2

We reiterate (5), the total cost to achieve an ϵ error for the local SGD case in Algorithm 1 as

$$C_{\text{SGD}} = \frac{c_{\text{SGD}}\overline{K}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{\sigma^2}{\tau}\right]. \tag{49}$$

To get $C_{\text{SGD}|\overline{K}>1} < C_{\text{SGD}|\overline{K}=1}$, we need

$$\frac{c_{\text{SGD}}\overline{K}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{\sigma^2}{\tau}\right] < \frac{c_{\text{SGD}}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M-N}{M-1}\right) + \frac{\sigma^2}{\tau}\right]$$
$$\Rightarrow (\overline{K}-1)\frac{\sigma^2}{\tau} < \gamma^2\left(\frac{M-N}{M-1} - \overline{K}\frac{M-N\overline{K}}{M-\overline{K}}\right)$$
$$= \gamma^2\frac{M(\overline{K}-1)\left(N\overline{K}-M+N-\frac{N\overline{K}}{M}\right)}{(M-1)(M-\overline{K})}.$$

The right-hand side is positive for $\overline{K} = M/N$. However, even for $\overline{K} = M/N - 1$, on the r.h.s. we notice that

$$N\overline{K} - M + N - \frac{N\overline{K}}{M} = M - N - M + N - \frac{N}{M}\left(\frac{M}{N} - 1\right) < 0.$$

Hence, except for $\overline{K} = M/N$, we can see that for $\gamma \geq A_1\sigma^2/\tau$ to achieve $C_{\text{SGD}|\overline{K}>1} > C_{\text{SGD}|\overline{K}=1}$ we need to have

$$\left(\frac{M}{N} - 1\right)\frac{\sigma^2}{\tau} < \frac{A_1\sigma^2}{\tau}\left(\frac{M-N}{M-1}\right) \tag{50}$$
$$\Rightarrow \frac{M-N}{N} < A_1\left(\frac{M-N}{M-1}\right) \tag{51}$$
$$\therefore \frac{M-1}{N} < A_1 \tag{52}$$

completing the proof.

B.5. Proofs on Intermediate Lemmas

Proof of Lemma 5.

$$\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\| \leq \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\mathbb{E}_k[\mathbf{d}_m^{(k,i)}]\right\|$$
$$= \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\mathbb{E}_k\left[\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\|$$
$$\leq \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\|. \tag{53}$$

Next, using Assumption 1,

$$\begin{aligned}
&\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \leq L\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\mathbf{w}_m^{(k,i-1,l)}-\mathbf{w}^{(k,i-1)}\right]\right\| \\
&\leq L\eta\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\sum_{l'=0}^{l-1}\nabla F_m(\mathbf{w}_m^{(k,i-1,l')},\xi_m^{(k,i-1,l')})\right]\right\| \\
&\leq L\eta\sum_{l=0}^{\tau-1}\sum_{l'=0}^{l-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l')})-\nabla F_m(\mathbf{w}^{(k,i-1)})+\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\leq L\eta\tau\sum_{l'=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l')})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\|+L\eta\sum_{l=0}^{\tau-1}\sum_{l'=0}^{l-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\Rightarrow(1-L\eta\tau)\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\leq \frac{L\eta\tau(\tau-1)}{2}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}^{(k,i-1)})-\nabla F_m(\mathbf{w}^{(k,0)})+\nabla F_m(\mathbf{w}^{(k,0)})\right]\right\| \\
&\leq \frac{L^2\eta\tau(\tau-1)}{2}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|+\frac{L\eta\tau(\tau-1)}{2}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|. \qquad (54)
\end{aligned}$$

Using η such that $\eta \leq \frac{1}{10L\tau}$, we get

$$\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \leq \frac{5L\eta\tau(\tau-1)}{9}\left[L\left\|\mathbb{E}_k[\mathbf{w}^{(k,i-1)}]-\mathbf{w}^{(k,0)}\right\|+\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|\right]. \qquad (55)$$

Bounding $\left\|\mathbb{E}_k[\mathbf{w}^{(k,i-1)}]-\mathbf{w}^{(k,0)}\right\|$. Analogous to (39) we have that

$$\left\|\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right\| = \left\|-\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}-\widetilde{\eta}\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)}+\widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,r)}\right\|,$$

where $\widetilde{\mathbf{r}}^{(k,r)} := \sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}$. Hence we have

$$\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right]\right\| \leq \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]\right\|+\widetilde{\eta}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{q}}^{(k,i)}\right]\right\|+\widetilde{\eta}^2\left\|\mathbb{E}_k\left[\widetilde{\mathbf{r}}^{(k,r)}\right]\right\| \qquad (56)$$

First, we bound the first term in the RHS of (56) as follows:

$$\begin{aligned}
&\frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]\right\| = \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&= \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&\leq \frac{\widetilde{\eta}N\overline{K}}{\tau M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&\leq \frac{\widetilde{\eta}N\overline{K}}{\tau M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left(\frac{5\eta L(\tau-1)\tau}{9}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\eta L^2(\tau-1)\tau}{9}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|\right) \qquad \text{(Using (55))} \\
&= \frac{5\widetilde{\eta}\eta LN\overline{K}(\tau-1)}{9M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\widetilde{\eta}\eta NL^2(\tau-1)}{9}\sum_{i=1}^{r}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|. \qquad (57)
\end{aligned}$$

We can bound the second term in the RHS of (56) using Lemma 3 to get

$$\widetilde{\eta}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{q}}^{(k,i)}\right]\right\| \leq r\widetilde{\eta}N\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \alpha\widetilde{\eta}Nr. \tag{58}$$

Third, we can bound the third term in the RHS of (56) as following:

$$\begin{aligned}
&\widetilde{\eta}^2\left\|\mathbb{E}_k\left[\widetilde{\mathbf{r}}^{(k,r)}\right]\right\| = \widetilde{\eta}^2\left\|\mathbb{E}_k\left[\sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right]\right\| \\
&\leq \widetilde{\eta}^2\sum_{i=1}^{r-1}\left\|\mathbb{E}_k\left[\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right]\right\|\left\|\mathbb{E}_k\left[\overline{\mathbf{S}}^{(k,i+1)}\right]\right\|\left\|\mathbb{E}_k\left[\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right]\right\| \\
&\leq \frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left\|\sum_{j'=1}^{i}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,j')}/\tau+\overline{\mathbf{q}}^{(k,j')}\right]\right\| && \text{(Using (30), (31) and } \widetilde{\eta}\leq\tfrac{1}{7LN\overline{K}}) \\
&\leq \frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left(\frac{5\eta LN\overline{K}(\tau-1)}{9M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| + \frac{5\eta NL^2(\tau-1)}{9}\sum_{j'=1}^{i}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right]\right\|\right) \\
&\quad + \frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left(iN\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + i\alpha N\right) && \text{(Using (57), (58))} \\
&\leq \frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\tau-1)(r-1)}{3M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| + \frac{3\widetilde{\eta}^2N^2Lr(r-1)\alpha}{5} \\
&\quad + \frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(r-1)}{3}\sum_{j'=1}^{r-1}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right]\right\| + \frac{3\widetilde{\eta}^2N^2Lr(r-1)}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|.
\end{aligned} \tag{59}$$

Plugging in (57), (58), and (59) into (56), and summing over r, we have

$$\begin{aligned}
\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right]\right\| &\leq \left(\frac{5\widetilde{\eta}\eta LN\overline{K}(\tau-1)}{9M}+\frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\tau-1)(r-1)}{3M}\right)\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| \\
&\quad + \left(\frac{5\widetilde{\eta}\eta NL^2(\tau-1)}{9}+\frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(r-1)}{3}\right)\sum_{i=1}^{r}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\| \\
&\quad + \left(r\widetilde{\eta}N+\frac{3\widetilde{\eta}^2N^2Lr(r-1)}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \alpha\widetilde{\eta}Nr + \frac{3\widetilde{\eta}^2N^2Lr(r-1)\alpha}{5} \\
&\quad \therefore \sum_{r=1}^{\overline{K}}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right]\right\| \\
&\leq \left(\frac{5\widetilde{\eta}\eta LN\overline{K}(\overline{K}-1)(\tau-1)}{9M}+\frac{\widetilde{\eta}^2\eta N^2L^2\overline{K}^2(\tau-1)(\overline{K}-1)}{3M}\right)\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| \\
&\quad + \left(\frac{5\widetilde{\eta}\eta NL^2\overline{K}(\tau-1)}{9}+\frac{\widetilde{\eta}^2\eta N^2L^3\overline{K}(\tau-1)(\overline{K}-1)}{3}\right)\sum_{i=1}^{\overline{K}}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\| \\
&\quad + \left(\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N}{2}+\frac{\widetilde{\eta}^2N^2L\overline{K}^2(\overline{K}-1)}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{49\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{100}
\end{aligned}$$

$$\sum_{r=1}^{\overline{K}} \left\| \mathbb{E}_k \left[\mathbf{w}^{(k,r-1)} - \mathbf{w}^{(k,0)} \right] \right\| \leq \frac{31\widetilde{\eta} N \overline{K}(\overline{K}-1)}{500M} \sum_{i=1}^{\overline{K}} \sum_{m \in \sigma(i)} \left\| \nabla F_m(\mathbf{w}^{(k,0)}) \right\| + \frac{27\widetilde{\eta} N \overline{K}^2}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|$$
$$+ \frac{\overline{K}(\overline{K}-1)\widetilde{\eta} N \alpha}{2} \quad \left(\because \frac{5\widetilde{\eta}\eta N L^2 \overline{K}(\tau-1)}{9} + \frac{\widetilde{\eta}^2 \eta N^2 L^3 \overline{K}(\tau-1)(\overline{K}-1)}{3} \leq 3/500 \right)$$
$$\sum_{r=1}^{\overline{K}} \left\| \mathbb{E}_k \left[\mathbf{w}^{(k,r-1)} - \mathbf{w}^{(k,0)} \right] \right\| \leq \frac{31\widetilde{\eta} N \overline{K}^2}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{31\widetilde{\eta} N \overline{K}(\overline{K}-1)\nu}{500} + \frac{\overline{K}(\overline{K}-1)\widetilde{\eta} N \alpha}{2} \tag{60}$$

$$\begin{aligned}
&\left\| \sum_{i=1}^{\overline{K}} \mathbb{E}_k \left[\sum_{m \in \mathcal{S}^{(k,i)}} \mathbf{d}_m^{(k,i)} \right] \right\| \\
&\leq \frac{N\overline{K}}{M} \sum_{i=1}^{\overline{K}} \sum_{m \in \sigma(i)} \left(\frac{5\eta L(\tau-1)\tau}{9} \left\| \nabla F_m(\mathbf{w}^{(k,0)}) \right\| + \frac{5\eta L^2(\tau-1)\tau}{9} \left\| \mathbb{E}_k \left[\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)} \right] \right\| \right) && \text{(Using (55))} \\
&\leq \frac{5\eta L N \overline{K}(\tau-1)\tau}{9} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{5\eta L N \overline{K}(\tau-1)\tau\nu}{9} \\
&\quad + \frac{5\eta N L^2(\tau-1)\tau}{9} \left(\frac{31\widetilde{\eta} N \overline{K}^2}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{31\widetilde{\eta} N \overline{K}(\overline{K}-1)\nu}{500} + \frac{\overline{K}(\overline{K}-1)\widetilde{\eta} N \alpha}{2} \right) && \text{(Using (60))} \\
&\leq \frac{3\eta L N \overline{K}(\tau-1)\tau}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{14\eta L N \overline{K}(\tau-1)\tau\nu}{25} + \frac{\eta L N(\overline{K}-1)(\tau-1)\tau\alpha}{36}.
\end{aligned}$$

□

Proof of Lemma 6. From (59) it follows that

$$\begin{aligned}
\widetilde{\eta}^2 \left\| \mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}] \right\| &\leq \frac{2\widetilde{\eta}^2 \eta N^2 L^2 \overline{K}(\overline{K}-1)(\tau-1)}{3M} \sum_{j'=1}^{\overline{K}} \sum_{m \in \sigma(j')} \left\| \nabla F_m(\mathbf{w}^{(k,0)}) \right\| + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \\
&\quad + \frac{2\widetilde{\eta}^2 \eta N^2 L^3(\tau-1)(\overline{K}-1)}{3} \sum_{j'=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\mathbf{w}^{(k,j'-1)} - \mathbf{w}^{(k,0)} \right] \right\| + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)\alpha}{5} \\
&\leq \frac{2\widetilde{\eta}^2 \eta N^2 L^2 \overline{K}(\overline{K}-1)(\tau-1)}{3M} \sum_{j'=1}^{\overline{K}} \sum_{m \in \sigma(j')} \left\| \nabla F_m(\mathbf{w}^{(k,0)}) \right\| + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \\
&\quad + \frac{2\widetilde{\eta}^2 \eta N^2 L^3(\tau-1)(\overline{K}-1)}{3} \left(\frac{31\widetilde{\eta} N \overline{K}^2}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{31\widetilde{\eta} N \overline{K}^2 \nu}{500} + \frac{\overline{K}(\overline{K}-1)\widetilde{\eta} N \alpha}{2} \right) && \text{(Using (60))} \\
&\quad + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)\alpha}{5} \\
&\leq \frac{2\widetilde{\eta}^2 \eta N^2 L^2 \overline{K}(\overline{K}-1)(\tau-1)}{3} \left[\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \nu \right] + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| \\
&\qquad \text{(Recall } \nu = \gamma + \alpha\text{, Assumption 3)} \\
&\quad + \frac{2\widetilde{\eta}^2 \eta N^2 L^3(\tau-1)(\overline{K}-1)}{3} \left(\frac{31\widetilde{\eta} N \overline{K}^2}{50} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \frac{31\widetilde{\eta} N \overline{K}^2 \nu}{500} + \frac{\overline{K}(\overline{K}-1)\widetilde{\eta} N \alpha}{2} \right) \\
&\quad + \frac{3\widetilde{\eta}^2 N^2 L \overline{K}(\overline{K}-1)\alpha}{5} \\
&\leq \widetilde{\eta} N(\overline{K}-1) \left(\frac{2\widetilde{\eta}^2 N L^2 \overline{K}}{3} + \frac{3\widetilde{\eta} N L \overline{K}}{5} + \frac{31\widetilde{\eta}^3 N^2 L^3 \overline{K}^2}{75} \right) \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|
\end{aligned}$$

$$+ \frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)}{3}\left(1+\frac{31\widetilde{\eta}NL\overline{K}}{500}\right)\nu + \widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\left(\frac{\widetilde{\eta}^2NL^2\overline{K}}{3}+\frac{3}{5}\right)\alpha$$
$$\leq \frac{17\widetilde{\eta}N\overline{K}}{250}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{84\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)}{125}\nu + \frac{16\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)}{25}\alpha.$$

□

Proof of Lemma 7. We have

$$\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$
$$= \frac{L}{2}\mathbb{E}_k\left[\left\|-\widetilde{\eta}\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}/\tau - \widetilde{\eta}\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)} + \widetilde{\eta}^2\underbrace{\left(\sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right)}_{:=\widetilde{\mathbf{r}}^{(k,r)}}\right\|^2\right]$$
$$\leq \frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right] + \frac{3L\widetilde{\eta}^2}{2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right] + \frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\widetilde{\mathbf{r}}^{(k,r)}\right\|^2\right]. \tag{61}$$

We first bound the last term in (61) as the following:

$$\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\widetilde{\mathbf{r}}^{(k,r)}\right\|^2\right] = \frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]$$
$$\leq \frac{54N^2L^3\widetilde{\eta}^4(r-1)}{25}\sum_{i=1}^{r-1}\mathbb{E}_k\left[\frac{2}{\tau^2}\left\|\sum_{j'=1}^{i}\overline{\mathbf{d}}^{(k,j')}\right\|^2 + 2\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right], \tag{62}$$

which follows from (30), (31). Note that the two terms in (62) are in similar forms as in the first two terms in (61). We will come back to bounding (62) after bounding the first two terms in (61).

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right] = \mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right] \leq r\sum_{i=1}^{r}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]$$
$$\leq r\sum_{i=1}^{r}\mathbb{E}_k\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}) + \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)})\right)\right\|^2$$
$$\leq \frac{rN^2\tau\overline{K}}{M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200} + \frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\right.$$
$$\left. + \frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\right) + r^2N\tau\sigma^2 \qquad \text{(Using the bound in (68))}$$
$$= \frac{69rN^2\tau^3(\tau-1)\overline{K}\eta^2L^2}{50M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 + r^2N\tau\sigma^2\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right)$$
$$+ \frac{69rN^2\tau^3(\tau-1)\overline{K}\eta^2L^4}{50M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right] \tag{63}$$

For the second term in (61), similar to how we got (36), we have that

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right] \leq \frac{1}{3}Nr^2\left[2N\left(\alpha^2+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right) + \gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right]. \tag{64}$$

We can use (63) and (64) to bound (62) as following:

$$
\begin{aligned}
&\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\widetilde{\mathbf{r}}^{(k,r)}\right\|^2\right] \\
&\leq \frac{54N^2L^3\widetilde{\eta}^4(r-1)}{25}\sum_{i=1}^{r-1}\Bigg(\frac{69iN^2\tau(\tau-1)\overline{K}\eta^2L^2}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad+\frac{69iN^2\tau(\tau-1)\overline{K}\eta^2L^4}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{2i^2N\sigma^2}{\tau}\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right) \\
&\quad+8N^2i^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+4Ni^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+8N^2i^2\alpha^2\Bigg) \\
&\leq \frac{3N^4L^5\tau(\tau-1)\overline{K}\eta^2\widetilde{\eta}^4r(r-1)^2}{M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{29N^4L^3\widetilde{\eta}^4(r-1)^2r^2}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad+\frac{3N^4L^7\tau(\tau-1)\overline{K}\eta^2\widetilde{\eta}^4(r-1)^2r}{M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right] \\
&\quad+\frac{72N^3L^3\widetilde{\eta}^4(r-1)^2r^2\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{36N^3L^3\widetilde{\eta}^4(r-1)^2r^2\sigma^2}{25\tau}\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right) \\
&\quad+\frac{29N^4L^3\widetilde{\eta}^4(r-1)^2r^2\alpha^2}{5}. \qquad (65)
\end{aligned}
$$

Finally, plugging in (63), (64), and (65) to (61) we have

$$
\begin{aligned}
&\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right\|^2\right] \\
&\leq \frac{2\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}r}{M}\left(2+3\widetilde{\eta}^2L^2N^2(r-1)^2\right)\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad+\frac{2\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}r}{M}\left(2+3\widetilde{\eta}^2L^2N^2(r-1)^2\right)\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right] \\
&\quad+\frac{2\widetilde{\eta}^2Nr^2\sigma^2}{\tau}\left(1+\frac{103\eta^2N\tau(\tau-1)L^2}{200}\right)\left(\frac{3}{2}+\frac{36\widetilde{\eta}^2L^2N^2(r-1)^2}{25}\right) \\
&\quad+2\widetilde{\eta}^2Nr^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(3+\frac{72\widetilde{\eta}^2L^2N^2(r-1)^2}{25}\right) \\
&\quad+\widetilde{\eta}^2N^2r^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2(r-1)^2}{5}\right)+\widetilde{\eta}^2N^2r^2\alpha^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2(r-1)^2}{5}\right) \\
\Rightarrow&\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\leq\frac{2\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}(r-1)}{M}\left(2+3\widetilde{\eta}^2L^2N^2\overline{K}^2\right)\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad+\frac{2\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}(r-1)}{M}\left(2+3\widetilde{\eta}^2L^2N^2\overline{K}^2\right)\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right] \\
&\quad+\frac{2\widetilde{\eta}^2N(r-1)^2\sigma^2}{\tau}\left(1+\frac{103\eta^2N\tau(\tau-1)L^2}{200}\right)\left(\frac{3}{2}+\frac{36\widetilde{\eta}^2L^2N^2\overline{K}^2}{25}\right) \\
&\quad+2\widetilde{\eta}^2N(r-1)^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(3+\frac{72\widetilde{\eta}^2L^2N^2\overline{K}^2}{25}\right) \\
&\quad+\widetilde{\eta}^2N^2r^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2r^2}{5}\right)+\widetilde{\eta}^2N^2(r-1)^2\alpha^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2r^2}{5}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{102\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}(r-1)}{25M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{61\widetilde{\eta}^2N^2(r-1)^2}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{102\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)(r-1)}{25}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{61\widetilde{\eta}^2N(r-1)^2\sigma^2}{20\tau}\\
&\quad+\frac{303\widetilde{\eta}^2N(r-1)^2\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2(r-1)^2\alpha^2}{5}\\
\Rightarrow&\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\leq\frac{51\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1}{50M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{15}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}^2}{50}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{61\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{60\tau}+\frac{101\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{15}\\
&\left(1-\frac{51\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}^2}{50}\right)\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\leq\frac{51\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1)}{50M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{15}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{61\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{60\tau}+\frac{101\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{15}
\end{aligned}
\tag{66}
$$

$$
\begin{aligned}
&\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\leq\frac{103\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1)}{100M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}\\
&\quad+\frac{203\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{100}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}\\
&\leq\frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}\\
&\quad+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}.
\end{aligned}
$$

□

Proof of Lemma 8.

$$
\begin{aligned}
&\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]=\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]\leq\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]\\
&=\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2
\end{aligned}
$$

$$= \frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)}))\right\|^2\right] \quad \text{(cross terms are zero)}$$

$$+\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)}))\right\|^2\right]$$

(independence of stochastic gradients across clients)

$$+\frac{3L\widetilde{\eta}^2\overline{K}N}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)})\right\|^2\right]$$

(unbiasedness of stochastic gradients)

$$+\frac{3L\widetilde{\eta}^2\overline{K}N}{2\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2M\tau^2}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\sigma^2+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2\tau}\sigma^2+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right] \tag{67}$$

From (67) we have that

$$\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$
$$\leq L^2\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\mathbf{w}_m^{(k,i-1,l)}-\mathbf{w}^{(k,i-1)})\right\|^2\right]$$
$$\leq \frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}+\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]. \tag{68}$$

Plugging in (68) to (67) we have

$$\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]$$
$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2\tau}\sigma^2+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\Bigg(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}$$
$$\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\Bigg)$$
$$=\frac{8L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{5\tau}+\frac{52L^5\widetilde{\eta}^2\eta^2\overline{K}N^2\tau(\tau-1)}{25}\sum_{i=1}^{\overline{K}}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$

$$
\begin{aligned}
&+\frac{52}{25}\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)\\
&=\frac{8L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{5\tau}+\frac{52L^5\widetilde{\eta}^2\eta^2\overline{K}N^2\tau(\tau-1)}{25}\left(\frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}\right.\\
&\left.+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}\right)\\
&+\frac{52}{25}\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)\\
&=\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\left(\frac{83\times 52\widetilde{\eta}^2L^2N^2\overline{K}^2}{25\times 20}+\frac{104}{25}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&+\frac{L\widetilde{\eta}^2N\overline{K}^2\sigma^2}{\tau}\left(\frac{8}{5}+\frac{26\times 51L^4\widetilde{\eta}^2\eta^2\overline{K}^2N^2\tau(\tau-1)}{25^2}\right)+\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\left(\frac{104}{25}+\frac{156\eta^2L^2\tau(\tau-1)}{2500}\right)\nu^2\\
&+\frac{52\times 51L^5\widetilde{\eta}^4\eta^2\overline{K}^4N^3\tau(\tau-1)\gamma^2}{25^2}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{52\times 41L^5\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4\tau(\tau-1)\alpha^2}{250}\\
&\le\frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau}+\frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}+\frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100}.
\end{aligned}
$$

□

Proof of Lemma 9.

$$
\begin{aligned}
\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right]&=\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&\le\frac{3L\widetilde{\eta}^4\overline{K}}{2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}\left[\left\|\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&\le\frac{54\widetilde{\eta}^4\overline{K}N^2L^3}{25}\sum_{i=1}^{\overline{K}-1}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right],
\end{aligned}
\tag{69}
$$

where (69) is derived in the same was as (35). Now we bound the following in (69)

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]&=\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{d}}^{(k,j')}/\tau+\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\le 2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{d}}^{(k,j')}/\tau\right\|^2\right]+2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&=\frac{2}{\tau^2}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]+2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right].
\end{aligned}
\tag{70}
$$

We have that

$$
\frac{2}{\tau^2}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]\le\frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]
$$

$$
\begin{aligned}
&= \frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]\\
&= \frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})\right\|^2\right]\\
&\quad+\frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]\\
&\le \frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})\right\|^2\right]\\
&\quad+\frac{2iN}{\tau}\sum_{j'=1}^{i}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]\\
&\le \frac{2iN\overline{K}}{M\tau^2}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\sum_{l=0}^{\tau-1}\sigma^2+\frac{2iN^2\overline{K}}{M\tau}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\sum_{l=0}^{\tau-1}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{2iN^2\overline{K}}{M\tau}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}+\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\right.\\
&\quad\left.+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\right)\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\quad+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\sum_{j'=1}^{\overline{K}}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25}(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)\\
&\quad+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\left(\frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}\right.\\
&\quad\left.+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}\right)\\
&\le i^2N\sigma^2\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+iN^2\eta^2L^2\tau(\tau-1)\overline{K}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(\frac{138}{25}+\frac{69\times 83L^2\widetilde{\eta}^2N^2\overline{K}^2}{1250}\right)\\
&\quad+\frac{69\times 51iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\sigma^2}{1250}+iN^2\eta^2L^2\tau(\tau-1)\overline{K}\nu^2\left(\frac{138}{25}+\frac{207L^2\eta^2\tau(\tau-1)}{2500}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{69 \times 51 i N^3 \eta^2 L^4 (\tau-1) \widetilde{\eta}^2 \overline{K}^3 \gamma^2}{25^2} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) + \frac{69 \times 41 i \widetilde{\eta}^2 \eta^2 N^4 L^4 \tau (\tau-1) \overline{K}^2 (\overline{K}-1) \alpha^2}{250} \\
&\leq i^2 N \sigma^2 \left(\frac{2}{\tau} + \frac{103 N \eta^2 L^2 (\tau-1)}{100} \right) + \frac{28 i N^2 \eta^2 L^2 \tau (\tau-1) \overline{K}}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{141 i N^3 \eta^2 L^4 (\tau-1) \widetilde{\eta}^2 \overline{K}^3 \sigma^2}{50} \\
&+ \frac{111 i N^2 \eta^2 L^2 \tau (\tau-1) \overline{K} \nu^2}{20} + \frac{113 i N^3 \eta^2 L^4 (\tau-1) \widetilde{\eta}^2 \overline{K}^3 \gamma^2}{20} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \\
&+ \frac{23 i \widetilde{\eta}^2 \eta^2 N^4 L^4 \tau (\tau-1) \overline{K}^2 (\overline{K}-1) \alpha^2}{2}. \qquad (71)
\end{aligned}
$$

Plugging in (71) to (70) in (69) we have

$$
\begin{aligned}
&\frac{3L\widetilde{\eta}^4}{2} \mathbb{E} \left[\left\| \widetilde{\mathbf{r}}^{(k,0)} \right\|^2 \right] \\
&\leq \frac{54 \widetilde{\eta}^4 \overline{K} N^2 L^3}{25} \sum_{i=1}^{\overline{K}-1} \left(i^2 N \sigma^2 \left(\frac{2}{\tau} + \frac{103 N \eta^2 L^2 (\tau-1)}{100} \right) + \frac{28 i N^2 \eta^2 L^2 \tau (\tau-1) \overline{K}}{5} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 \right. \\
&+ \frac{141 i N^3 \eta^2 L^4 (\tau-1) \widetilde{\eta}^2 \overline{K}^3 \sigma^2}{50} + \frac{111 i N^2 \eta^2 L^2 \tau (\tau-1) \overline{K} \nu^2}{20} + \frac{113 i N^3 \eta^2 L^4 (\tau-1) \widetilde{\eta}^2 \overline{K}^3 \gamma^2}{20} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \\
&\left. + \frac{23 i \widetilde{\eta}^2 \eta^2 N^4 L^4 \tau (\tau-1) \overline{K}^2 (\overline{K}-1) \alpha^2}{2} + 2\mathbb{E} \left[\left\| \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right\|^2 \right] \right) \\
&\leq \frac{18 \widetilde{\eta}^4 \overline{K}^3 (\overline{K}-1) N^3 L^3 \sigma^2}{25} \left(\frac{2}{\tau} + \frac{103 N \eta^2 L^2 (\tau-1)}{100} \right) + \frac{54 \times 14 \widetilde{\eta}^4 \eta^2 \overline{K}^3 (\overline{K}-1) N^4 L^5 (\tau-1) \tau}{125} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 \\
&+ \frac{27 \times 141 \widetilde{\eta}^6 \eta^2 \overline{K}^5 (\overline{K}-1) N^5 L^7 (\tau-1) \sigma^2}{1250} + \frac{27 \times 111 \widetilde{\eta}^4 \eta^2 \overline{K}^3 (\overline{K}-1) N^4 L^5 (\tau-1) \tau \nu^2}{500} \\
&+ \frac{27 \times 113 \widetilde{\eta}^6 \eta^2 \overline{K}^5 (\overline{K}-1) N^5 L^7 (\tau-1) \gamma^2}{500} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \\
&+ \frac{54 \times 23 \widetilde{\eta}^6 \eta^2 N^6 L^7 \tau (\tau-1) \overline{K}^4 (\overline{K}-1)^2 \alpha^2}{100} + \frac{108 \widetilde{\eta}^4 \overline{K} N^2 L^3}{25} \sum_{i=1}^{\overline{K}-1} \mathbb{E} \left[\left\| \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right\|^2 \right] \\
&\leq \frac{18 \widetilde{\eta}^4 \overline{K}^3 (\overline{K}-1) N^3 L^3 \sigma^2}{25} \left(\frac{2}{\tau} + \frac{103 N \eta^2 L^2 (\tau-1)}{100} \right) + \frac{54 \times 14 \widetilde{\eta}^4 \eta^2 \overline{K}^3 (\overline{K}-1) N^4 L^5 (\tau-1) \tau}{125} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 \\
&+ \frac{27 \times 141 \widetilde{\eta}^6 \eta^2 \overline{K}^5 (\overline{K}-1) N^5 L^7 (\tau-1) \sigma^2}{1250} + \frac{27 \times 111 \widetilde{\eta}^4 \eta^2 \overline{K}^3 (\overline{K}-1) N^4 L^5 (\tau-1) \tau \nu^2}{500} \\
&+ \frac{27 \times 113 \widetilde{\eta}^6 \eta^2 \overline{K}^5 (\overline{K}-1) N^5 L^7 (\tau-1) \gamma^2}{500} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) + \frac{54 \times 23 \widetilde{\eta}^6 \eta^2 N^6 L^7 \tau (\tau-1) \overline{K}^4 (\overline{K}-1)^2 \alpha^2}{100} \\
&+ \frac{108 \widetilde{\eta}^4 \overline{K} N^2 L^3}{25} \left(2\overline{K}^3 N \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + 4 N^2 \overline{K}^3 \left\| \nabla F(\mathbf{w}^{(k,0)} \right\|^2 + \frac{4 N^2 \overline{K}^2 (\overline{K}-1) \alpha^2}{3} \right) \\
&\leq \widetilde{\eta} N \overline{K} (\widetilde{\eta}^3 \overline{K}^3 N^3 L^3) \left(\frac{108 \times 4}{25} + \frac{54 \times 14 \widetilde{\eta}^2 L^2}{125} \right) \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{27 \times 141 \widetilde{\eta}^2 L \overline{K}^2 N \sigma^2}{1250 \tau} (\widetilde{\eta}^4 \eta^2 \overline{K}^4 N^4 L^6 \tau (\tau-1)) \\
&+ \frac{18 \widetilde{\eta}^2 L \overline{K}^2 N \sigma^2}{25 \tau} \left(2 \widetilde{\eta}^2 \overline{K}^2 N^2 L^2 + \frac{103 \widetilde{\eta}^2 \overline{K}^2 N^3 L^4 \eta^2 \tau (\tau-1)}{100} \right) \\
&+ \widetilde{\eta}^4 \overline{K}^4 N^3 L^3 \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 \left(\frac{216}{25} + \frac{27 \times 113 \widetilde{\eta}^2 \eta^2 \overline{K}^2 N^3 L^4 (\tau-1)}{500} \right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{27 \times 111 \widetilde{\eta}^2 \eta^2 L^3 \overline{K}^2 N^2 \tau(\tau-1)\nu^2}{500} (\widetilde{\eta}^2 \overline{K}^2 N^2 L^2) \\
&+ \widetilde{\eta}^4 \overline{K}^3 (\overline{K}-1) N^4 L^3 \alpha^2 \left(\frac{108 \times 4}{75} + \frac{54 \times 23 \widetilde{\eta}^2 \eta^2 \overline{K}^2 N^2 L^4 \tau(\tau-1)}{100} \right) \\
&\leq \frac{9\widetilde{\eta} N \overline{K}}{500} \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \frac{3\widetilde{\eta}^2 L \overline{K}^2 N \sigma^2}{200\tau} + \frac{173 \widetilde{\eta}^4 \overline{K}^4 N^3 L^3 \gamma^2}{20} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \\
&+ \frac{3\widetilde{\eta}^2 \eta^2 L^3 \overline{K}^2 N^2 \tau(\tau-1)\nu^2}{50} + \frac{59 \widetilde{\eta}^4 \overline{K}^3 (\overline{K}-1) N^4 L^3 \alpha^2}{10}
\end{aligned}
\tag{72}
$$

□

C. Proofs for the CyCP+Shuffled SGD Case (Case (iii) in Algorithm 1)

Now let us extend our analysis to clients locally performing shuffled SGD. Recall that for shuffled SGD, as shown in Algorithm 1 we have that each client in $m \in \mathcal{S}^{(k,i)}$ receives the global model $\mathbf{w}^{(k,i-1)}$ and initializes its local model as the global model i.e., $\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$. Then the client performs shuffled SGD over its B components sequentially with update rule $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1,l)}),\ l \in [0, ..., B-1]$ where $\pi_m^k \sim \text{Unif}(\mathcal{P}_B)$ and $F_{m,\pi_m^k(l)}(\mathbf{w})$ is the $\pi_m^k(l)^{\text{th}}$ component of the local loss of client m such that the sum of all the components for each client is equal to the local loss of that client i.e., $F_m(\mathbf{w}) = \frac{1}{B}\sum_{l=0}^{B-1} F_{m,l}(\mathbf{w})$. Hence the update rule over the inner loop $i \in [\overline{K}]$ is as follows:

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}) \tag{73}$$

Using

$$\begin{aligned}
&\nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}) = \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}) - \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) \\
&= \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \underbrace{\int_0^1 \nabla^2 F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)} + t(\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)}))dt}_{:=\widehat{\mathbf{H}}_{m,l}^{(k,i-1)}}(\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)})
\end{aligned} \tag{74}$$

we have that

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left[\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \widehat{\mathbf{H}}_{m,l}^{(k,i-1)}(\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)}) \right]. \tag{75}$$

Leveraging the fact that $\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$ we can use recursion to get the update rule

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) \tag{76}$$

Similarly, we can define

$$\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) = \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \underbrace{\int_0^1 \nabla^2 F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)} + t(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})dt}_{:=\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}),$$

to get

$$\begin{aligned}
\mathbf{w}^{(k,i)} &= \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \left(\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}) \right) \\
&= \mathbf{w}^{(k,i-1)} - \eta \times \underbrace{\frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})}_{:=\mathbf{t}^{(k,i)}} \\
&\quad - \eta \times \underbrace{\frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}}_{:=\mathbf{T}^{(k,i)}} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})
\end{aligned} \tag{77}$$

Unrolling (77) we can obtain the update rule for the outer loop as follows:

$$\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,0)} - \eta \sum_{i=1}^{\overline{K}} \left(\prod_{j=\overline{K}}^{i+1} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{t}^{(k,i)} \tag{78}$$

Applying summation by parts to $\prod_{j=\overline{K}}^{i+1}(\mathbf{I}_d - \eta\mathbf{T}^{(k,j)})$ and $\mathbf{t}^{(k,i)}$ in (78) and then again to $\prod_{j=B-1}^{l+1}\left(\mathbf{I} - \eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)$ and $\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})$ in $\mathbf{t}^{(k,i)}$ we can rewrite (78) as

$$
\begin{aligned}
\mathbf{w}^{(k+1)} &= \mathbf{w}^{(k,0)} - \frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \\
&+ \eta^2 \times \underbrace{\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d - \eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})}_{:=\mathbf{r}_1^{(k,0)}} \\
&+ \eta^2 \times \underbrace{\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d - \eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)}_{:=\mathbf{r}_2^{(k,0)}} \\
&\underbrace{\frac{-\eta^3}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d - \eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d - \eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right)}_{:=\mathbf{r}_3^{(k,0)}} \\
&= \mathbf{w}^{(k,0)} - \frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \eta^2\mathbf{r}_1^{(k,0)} + \eta^2\mathbf{r}_2^{(k,0)} - \eta^3\mathbf{r}_3^{(k,0)} \qquad (79)
\end{aligned}
$$

Taking expectation conditioned on all the past till $\mathbf{w}^{(k,0)}$, we get

$$
\mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} = -\eta\overline{K}B\nabla F(\mathbf{w}^{(k,0)}) + \eta^2\mathbb{E}_k[\underbrace{\mathbf{r}_1^{(k,0)} + \mathbf{r}_2^{(k,0)} - \eta\mathbf{r}_3^{(k,0)}}_{:=\mathbf{r}^{(k,0)}}] \qquad (80)
$$

which follows from the expectation over the selected client set $\mathcal{S}^{(k,i)},\ i \in [\overline{K}]$.

C.1. Intermediate Results

Lemma 10.

$$
\begin{aligned}
\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\| &\le \left(\frac{2e^{1/10}}{3} + \frac{2exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{3}\right)\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}\overline{K}L(B^{3/2}-1)\overline{\nu} \\
&+ \left(\frac{e^{1/10}}{2} + \frac{exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{2}\right)\overline{K}LB(B-1)\nu + \frac{exp(e^{1/10}/10)e^{1/10}B^2L\overline{K}(\overline{K}-1)\alpha}{2} \\
&+ \left(\frac{e^{1/10}}{2} + \frac{exp(e^{1/10}/10)e^{1/10}(\overline{K}-1)}{2} + \frac{exp(e^{1/10}/10)e^{1/5}\eta(B-1)L(\overline{K}-1)}{2}\right)\overline{K}LB^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|.
\end{aligned}
$$

Lemma 11.

$$
\begin{aligned}
&3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2 + \left\|\mathbf{r}_2^{(k,0)}\right\|^2 + \left\|\mathbf{r}_3^{(k,0)}\right\|^2\right] \\
&\le \frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log(4MBK/\delta)\overline{\nu}^2 + \frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2} + \frac{3\eta\overline{K}B}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 \\
&+ \frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2exp(e^{1/10}/5)e^{1/5}.
\end{aligned}
$$

C.2. Proof for Theorem 3

Theorem. *With Assumption 1-Assumption 3, and Assumption 5 we have that with step-size* $\eta = \log(MBK^2)/\mu\overline{K}BK$ *and* $K \geq 10\kappa\log(MBK^2)$ *where* $\kappa = L/\mu$ *we have that the convergence error is bounded as:*

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq (1-\mu\eta\overline{K}B)^K(F(\mathbf{w}^{(0,0)}) - F^*) + \mathcal{O}\left(\frac{\kappa\eta B\overline{K}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\right)\\ + \tilde{\mathcal{O}}\left(\eta^2(B-1)\kappa L\overline{\nu}^2 + \eta^2\kappa L(B-1)^2\nu^2\right)\\ + \mathcal{O}\left((\overline{K}-1)^2B^2L\kappa\eta^2\alpha^2\right).\end{aligned}$$

With $\eta = \log(MBK^2)/\mu B\overline{K}K$*, we get*

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MBK^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) + \widetilde{\mathcal{O}}\left(\frac{(B-1)}{B^2}\frac{\kappa^2\overline{\nu}^2}{\mu\overline{K}^2K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2\overline{K}^2K^2}\right)\\ + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right),\end{aligned}$$

where $\nu = \gamma + \alpha$ *and* $\tilde{\mathcal{O}}(\cdot)$ *subsumes all log-terms and constants.*

Corollary 2. *Recall that the total communication rounds translates to* $T = \overline{K}K$ *and therefore in terms of* T *we have that the bound becomes*

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MBT^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right)\\ + \widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{\nu}^2}{\mu T^2}\frac{(B-1)}{B^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa\overline{K}\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right).\end{aligned}$$

Since $K \geq 10\kappa\log(MBK^2)$, we have $T \geq 10\kappa\overline{K}\log(MB\overline{K}^2T^2)$ and one cannot increase $\overline{K}$ without increasing T accordingly due to its lower bound depending on $\overline{K}$.

Proof. Using the L-smoothness property of the global objective F we have

$$\begin{aligned}&\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq \langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle + \frac{L}{2}\mathbb{E}\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right]\\ &\leq -\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \eta^2\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\|\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right].\end{aligned} \tag{81}$$

Plugging Lemma 10, we bound the second term in the RHS of (81).

$$\begin{aligned}&\eta^2\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\|\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\ &\leq \left(\frac{2e^{1/10}}{3} + \frac{2\exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{3}\right)\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\ &\quad + \left(\frac{e^{1/10}}{2} + \frac{\exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{2}\right)\eta^2\overline{K}LB(B-1)\nu\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\ &\quad + \frac{\eta^2\exp(e^{1/10}/10)e^{1/10}B^2L\overline{K}(\overline{K}-1)\alpha}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\ &\quad + \left(\frac{e^{1/10}}{2} + \frac{\exp(e^{1/10}/10)e^{1/10}(\overline{K}-1)}{2} + \frac{\exp(e^{1/10}/10)e^{1/5}\eta(B-1)L(\overline{K}-1)}{2}\right)\eta^2\overline{K}LB^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\ &\leq -\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{47}{20}\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{63\eta^2\overline{K}LB(B-1)\nu}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| \\
&+ \frac{31\eta^2B^2L\overline{K}(\overline{K}-1)\alpha}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| \\
&+ \left(\frac{e^{1/10}}{20} + \frac{\exp(e^{1/10}/10)e^{1/10}}{20} + \frac{\exp(e^{1/10}/10)e^{1/5}}{200}\right)\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 \\
\leq\ & \frac{13}{100}\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{47}{20}\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| \\
&+ \frac{63\eta^2\overline{K}LB(B-1)\nu}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \frac{31\eta^2B^2L\overline{K}(\overline{K}-1)\alpha}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|. \qquad (82)
\end{aligned}
$$

Next we bound the third term in the RHS of (81) as follows:

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right] = \frac{L}{2}\mathbb{E}_k\left[\left\|-\frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \eta^2\mathbf{r}^{(k,0)}\right\|^2\right] \\
&\leq L\mathbb{E}_k\left[\left\|-\frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right] + L\mathbb{E}_k\left[\left\|\eta^2\mathbf{r}^{(k,0)}\right\|^2\right] \\
&\leq L\eta^2B^2\overline{K}^2\mathbb{E}_k\left[\left\|\frac{1}{\overline{K}}\sum_{i=1}^{\overline{K}}\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right] + 3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2 + \left\|\mathbf{r}_2^{(k,0)}\right\|^2 + \eta^2\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right] \\
&\leq L\eta^2B^2\overline{K}^2\left[\frac{1}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + \left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right] + 3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2 + \left\|\mathbf{r}_2^{(k,0)}\right\|^2 + \eta^2\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]. \qquad (83)
\end{aligned}
$$

Plugging the bound from Lemma 11 in (83) we have

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right] \leq \frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + \frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2 \\
&+ \frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2} + \frac{\eta\overline{K}B}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{31\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2}{50}. \qquad (84)
\end{aligned}
$$

Finally, substituting (82) and (83) into (81) we get

$$
\begin{aligned}
&\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \\
&\leq -\frac{82}{100}\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \\
&+ \left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{47\eta^{3/2}\overline{K}^{1/2}L(B^{3/2}-1)\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\overline{\nu}}{2B^{1/2}}\right) \\
&+ \left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{63\eta^{3/2}\overline{K}^{1/2}LB^{1/2}(B-1)\nu}{10}\right) \\
&+ \left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{31\eta^{3/2}\overline{K}^{1/2}(\overline{K}-1)LB^{3/2}\alpha}{2}\right) \\
&+ \frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2 + \frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2} + \frac{31\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2}{50} \\
\Rightarrow &\mathbb{E}[F(\mathbf{w}^{(k+1,0)})] - F^* \leq (1-\mu\eta B\overline{K})(\mathbb{E}[F(\mathbf{w}^{(k,0)})] - F^*) + \frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2
\end{aligned}
$$

$$+\left(\frac{47^2(B^{3/2}-1)^2}{8B}+\frac{31(B-1)^2}{10}\right)\eta^3\overline{K}L^2\log\left(\frac{4MBK}{\delta}\right)\overline{\nu}^2+\left(\frac{63^2}{200}+\frac{1}{2}\right)\eta^3\overline{K}L^2B(B-1)^2\nu^2$$
$$+121\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2. \tag{85}$$

Unrolling (85) we have

$$\mathbb{E}[F(\mathbf{w}^{(k+1,0)})]-F^*\leq(1-\mu\eta\overline{K}B)^K(F(\mathbf{w}^{(0,0)})-F^*)+\frac{21\kappa\eta B\overline{K}}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$
$$+\left(\frac{47^2(B^{3/2}-1)^2}{8B^2}+\frac{31(B-1)^2}{10B}\right)\eta^2\kappa L\log\left(\frac{4MBK}{\delta}\right)\overline{\nu}^2+\left(\frac{63^2}{200}+\frac{1}{2}\right)\eta^2\kappa L(B-1)^2\nu^2 \tag{86}$$
$$+121(\overline{K}-1)^2B^2L\kappa\eta^2\alpha^2$$

With $\eta=\log(MBK^2)/\mu B\overline{K}K$, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\\
&\leq\frac{F(\mathbf{w}^{(0,0)})-F^*}{MBK^2}+\frac{21\kappa\log(MBK^2)}{10\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\\
&+\left(\frac{47^2(B^{3/2}-1)^2}{8B^4}+\frac{31(B-1)^2}{10B^3}\right)\frac{\kappa^2\log(4MBK/\delta)\log^2(MBK^2)\overline{\nu}^2}{\mu\overline{K}^2K^2}+\frac{21\kappa^2(B-1)^2\log^2(MBK^2)\nu^2}{10\mu B^2\overline{K}^2K^2}\\
&+\frac{121(\overline{K}-1)^2\kappa^2\log^2(MBK^2)\alpha^2}{\mu\overline{K}^2K^2}\\
&=\frac{F(\mathbf{w}^{(0,0)})-F^*}{MBK^2}+\widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)+\widetilde{\mathcal{O}}\left(\left(\frac{(B^{3/2}-1)^2}{B^4}+\frac{(B-1)^2}{B^3}\right)\frac{\kappa^2\overline{\nu}^2}{\mu\overline{K}^2K^2}\right)\\
&+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2\overline{K}^2K^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right),
\end{aligned}$$

which concludes the proof. □

C.3. Proof for Corollary 3

Recall that the total cost for CyCP+Shuffled SGD and LocalRR respectively is as follows:

$$C_{\text{SSGD}}(\epsilon)=\widetilde{\mathcal{O}}\left(\frac{c_{\text{SSGD}}\kappa}{\sqrt{\mu\epsilon}}\left(\frac{\overline{K}\sqrt{\mu}}{\kappa\sqrt{MB}}+\nu+\overline{K}\alpha+\frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{87}$$
$$C_{\text{LocalRR}}(\epsilon)=\widetilde{\mathcal{O}}\left(\frac{\overline{K}c_{\text{SSGD}}}{\sqrt{\epsilon}}\left(\frac{1}{\sqrt{MB}}+\nu+\frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{88}$$

For $C_{\text{SSGD}}(\epsilon)<C_{\text{LocalRR}}(\epsilon)$ to be true, we need to have Hence, equivalently, we need to have

$$(1-\overline{K})\nu+\overline{K}\alpha+(1-\overline{K})\frac{\overline{\nu}}{\sqrt{B}}<0.$$

Since we have that $\overline{K}=M/N$, we have

$$\left(\frac{N-M}{N}\right)\nu+\frac{M\alpha}{N}+\frac{(N-M)\overline{\nu}}{N\sqrt{B}}<0,\ (N-M)\nu+M\alpha+\frac{(N-M)\overline{\nu}}{\sqrt{B}}<0,\ \left(\alpha-\nu-\frac{\overline{\nu}}{\sqrt{B}}\right)M<-N\nu-\frac{N\overline{\nu}}{\sqrt{B}}$$
$$\left(-\gamma-\frac{\overline{\nu}}{\sqrt{B}}\right)M<-N\left(\alpha+\gamma+\frac{\overline{\nu}}{\sqrt{B}}\right),\ M>N\left(\alpha+\gamma+\frac{\overline{\nu}}{\sqrt{B}}\right)/\left(\gamma+\frac{\overline{\nu}}{\sqrt{B}}\right)=N\left(1+\frac{\alpha}{\gamma+\frac{\overline{\nu}}{\sqrt{B}}}\right),$$

completing the proof.

C.4. Proofs on Intermediate Lemmas

Proof of Lemma 10. In the following, we bound all three components of $\mathbf{r}^{(k,0)}$ separately as follows.

$$
\begin{aligned}
\left\|\mathbb{E}_k[\mathbf{r}_1^{(k,0)}]\right\| &= \left\|\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right]\right\| \\
&= \left\|\frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\mathbb{E}_k\left[\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right]\right\| \\
&\qquad\qquad (\because \text{ unbiased client sampling}) \\
&\le \frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \\
&\le \frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\right\|\left\|\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\right\|\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \\
&\le \frac{(1+\eta L)^B\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \qquad (\because \left\|\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\right\|\le L) \\
&\le \frac{e^{1/10}\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \qquad (\text{Using } \eta\le\tfrac{1}{10BL}) \\
&\le \frac{e^{1/10}\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left(\overline{\nu}\sqrt{8(l+1)\log\left(\frac{4MBK}{\delta}\right)}+(l+1)\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|\right) \\
&\qquad\qquad (\text{Lemma 8 in (Yun et al., 2022)}) \\
&\le \frac{2e^{1/10}\overline{K}L(B^{3/2}-1)\overline{\nu}}{3}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{e^{1/10}\overline{K}LB(B-1)}{2M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| \\
&\le \frac{2e^{1/10}\overline{K}L(B^{3/2}-1)\overline{\nu}}{3}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{e^{1/10}\overline{K}LB(B-1)\nu}{2}\left[\nu+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right]. \qquad (89)
\end{aligned}
$$

We can similarly bound the next noise term as

$$
\begin{aligned}
\left\|\mathbb{E}_k[\mathbf{r}_2^{(k,0)}]\right\| &= \left\|\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right]\right\| \\
&\le \frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right]\right\| \qquad (\text{Jensen's inequality}) \\
&\le \frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\right]\right\|\left\|\mathbb{E}_k\left[\mathbf{T}^{(k,i+1)}\right]\right\|\left\|\mathbb{E}_k\left[\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right]\right\| \qquad (90)
\end{aligned}
$$

Since

$$
\begin{aligned}
\left\|\mathbf{T}^{(k,i)}\right\| &= \left\|\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\left(\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)\right)\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}\right\| \\
&\le \frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\left\|\left(\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)\right)\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}\right\|
\end{aligned}
$$

$$\leq \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left\| \prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right) \right\| \left\| \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)} \right\|$$

$$\leq \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} (1+\eta L)^B L \leq e^{1/10} BL, \tag{91}$$

since $\left\| \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)} \right\| \leq L$. We can bound (90) as

$$\left\| \mathbb{E}_k[\mathbf{r}_2^{(k,0)}] \right\| \leq (1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\sum_{j=1}^{i} \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \right] \right\|$$

$$= (1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \frac{\overline{K}}{M} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \right\| \qquad (\because \text{ unbiased client sampling})$$

$$= \frac{(1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} B^2 L \overline{K}}{M} \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \sum_{m\in\sigma^{(j)}} \nabla F_m(\mathbf{w}^{(k,0)}) \right\|$$

$$\leq \exp(e^{1/10}/10) e^{1/10} B^2 L \sum_{i=1}^{\overline{K}-1} \left(i \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + i\alpha \right) \qquad (\text{using Lemma 3, and } \eta \leq \tfrac{1}{10LB\overline{K}})$$

$$\leq \exp(e^{1/10}/10) e^{1/10} B^2 L \frac{\overline{K}(\overline{K}-1)}{2} \left[\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \alpha \right]. \tag{92}$$

For the last noise term, we have

$$\left\| \mathbb{E}_k[\mathbf{r}_3^{(k,0)}] \right\| \leq$$

$$\left\| \mathbb{E}_k \frac{1}{N} \sum_{i=1}^{\overline{K}-1} \left(\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{T}^{(k,i+1)} \left(\sum_{j=1}^{i} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \right.\right.$$
$$\left.\left. \times \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right) \right\|^2$$

$$\leq \sum_{i=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right] \right\| \left\| \mathbb{E}_k \left[\mathbf{T}^{(k,i+1)} \right] \right\|$$
$$\times \left\| \mathbb{E}_k \left[\sum_{j=1}^{i} \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right] \right\|$$

$$\leq (1+\eta e^{1/10} BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \frac{\overline{K}}{M} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\|$$
(Using (91))

$$\leq \frac{\exp(e^{1/10}/10) e^{1/10} BL\overline{K}}{M} \sum_{i=1}^{\overline{K}-1} \sum_{j=1}^{i} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \mathbb{E}_k \left[\left\| \prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right\| \left\| \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \right\| \left\| \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\| \right]$$
(Using $\eta \leq \frac{1}{10BL\overline{K}}$)

$$\leq \frac{(1+\eta L)^B \exp(e^{1/10}/10) e^{1/10} BL^2 \overline{K}(\overline{K}-1)}{M} \sum_{j=1}^{\overline{K}} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \mathbb{E}_k \left\| \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\|$$
(Using $\left\| \widehat{\mathbf{H}}_{m,l+1}^{(k,j)} \right\| \leq L$, for all j)

$$\leq \frac{\exp(e^{1/10}/10)e^{1/5}BL^2\overline{K}(\overline{K}-1)}{M}\sum_{j=1}^{\overline{K}}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}\left(\overline{\nu}\sqrt{8(l+1)\log\left(\frac{4MBK}{\delta}\right)}+(l+1)\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|\right)$$

(Lemma 8 in (Yun et al., 2022))

$$\leq \frac{2\exp(e^{1/10}/10)e^{1/5}B(B^{3/2}-1)L^2\overline{K}(\overline{K}-1)}{3}\overline{\nu}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}$$

$$+\frac{\exp(e^{1/10}/10)e^{1/5}B^2(B-1)L^2\overline{K}(\overline{K}-1)}{2M}\sum_{j=1}^{\overline{K}}\sum_{m\in\sigma^{(j)}}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|$$

$$\leq \exp(e^{1/10}/10)e^{1/5}BL^2\overline{K}(\overline{K}-1)\left[\frac{2(B^{3/2}-1)}{3}\overline{\nu}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{B(B-1)\nu}{2}+\frac{B(B-1)}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right]. \tag{93}$$

Finally, using (89), (92), (93), in

$$\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\| \leq \left\|\mathbb{E}_k[\mathbf{r}_1^{(k,0)}]\right\| + \left\|\mathbb{E}_k[\mathbf{r}_2^{(k,0)}]\right\| + \eta\left\|\mathbb{E}_k[\mathbf{r}_3^{(k,0)}]\right\|$$

we get the final bound. □

Proof of Lemma 11. Similar to how we bounded the norm of the noise terms, we can bound the norm square of the noise terms as following

$$3\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2\right] = 3\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$\leq 3\overline{K}(B-1)\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$\leq \frac{3\overline{K}^2(B-1)}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma^{(i)}}\sum_{l=0}^{B-2}(1+\eta L)^{2B}L^2\mathbb{E}_k\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2$$

$$\leq \frac{6e^{1/5}\overline{K}^2L^2(B-1)}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma^{(i)}}\sum_{l=0}^{B-2}\left((l+1)(8\log(4MBK/\delta)\overline{\nu}^2+(l+1)^2\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right)$$

$$\leq 4e^{1/5}\overline{K}^2L^2B(B-1)^2\left[6\log(4MBK/\delta)\overline{\nu}^2+B\nu^2+B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right], \tag{94}$$

and

$$3\mathbb{E}_k\left[\left\|\mathbf{r}_2^{(k,0)}\right\|^2\right] = 3\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right\|^2\right]$$

$$\leq \frac{3(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right\|^2\left\|\mathbf{T}^{(k,i+1)}\right\|^2\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$\leq \frac{3(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[(1+\eta BLe^{1/10})^{2\overline{K}}(B^2L^2e^{1/5})\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$
\begin{aligned}
&\leq \frac{3(\overline{K}-1)B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\leq \frac{3(\overline{K}-1)B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N^2}\left(\frac{1}{3}(\overline{K}-1)\overline{K}^2N\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\alpha^2\right)\right]\right)\\
&= \frac{(\overline{K}-1)^2\overline{K}^2B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N}\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\alpha^2\right)\right]. \qquad (95)
\end{aligned}
$$

and

$$
\begin{aligned}
3\eta^2\mathbb{E}_k\left[\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right] &= 3\eta^2\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\right.\right.\\
&\quad\left.\left.\times\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right)\right\|^2\right]\\
&\leq \frac{3\eta^2(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left(\left\|\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right\|\left\|\mathbf{T}^{(k,i+1)}\right\|\right.\right.\\
&\quad\left.\left.\times\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|\right)^2\right]\\
&\leq \frac{3\eta^2(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}(1+\eta BLe^{1/10})^{2\overline{K}}(B^2L^2e^{1/5})\\
&\quad\times(iN(B-1))\sum_{j=1}^{i}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\leq \frac{3\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^2\exp(e^{1/10}/5)e^{1/5}}{M}\\
&\quad\times\sum_{i=1}^{\overline{K}-1}i\sum_{j=1}^{i}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}\mathbb{E}_k\left[\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\leq \frac{3\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^2\exp(e^{1/10}/5)e^{1/5}}{M}\\
&\quad\times\sum_{i=1}^{\overline{K}-1}i\sum_{j=1}^{i}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}(1+\eta L)^{2B}L^2\mathbb{E}_k\left[\left\|\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\leq \frac{6\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^4\exp(e^{1/10}/5)e^{2/5}}{M}\\
&\quad\times\sum_{i=1}^{\overline{K}-1}i\sum_{j=1}^{i}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}\left(8(l+1)\log(4MBK/\delta)\overline{\nu}^2+(l+1)^2\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right)\\
&\leq 12\eta^2(\overline{K}-1)^2\overline{K}^2B^3(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}\log(4MBK/\delta)\overline{\nu}^2\\
&\quad+\frac{\eta^2(\overline{K}-1)^2\overline{K}^2B^4(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}}{M}\sum_{j=1}^{\overline{K}}\sum_{m\in\sigma^{(j)}}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2
\end{aligned}
$$

$$\leq 2\eta^2(\overline{K}-1)^2\overline{K}^2B^3(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}\left[6\log\left(4MBK/\delta\right)\overline{\nu}^2+B\left(\nu^2+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right)\right]. \tag{96}$$

Using (94), (95), and (96) we have

$$\begin{aligned}
&L\eta^4\left(3\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2\right]+3\mathbb{E}_k\left[\left\|\mathbf{r}_2^{(k,0)}\right\|^2\right]+3\eta^2\mathbb{E}_k\left[\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]\right)\\
&\leq\left(24e^{1/5}\eta\overline{K}LB+\frac{2\eta\overline{K}BL\exp(e^{1/10}/5)e^{2/5}}{25}\right)\eta^3\overline{K}L^2(B-1)^2\log\left(4MBK/\delta\right)\overline{\nu}^2\\
&\quad+\left(4e^{1/5}\eta\overline{K}LB+\frac{\eta(\overline{K}-1)BL\exp(e^{1/10}/5)e^{2/5}}{50}\right)\eta^3\overline{K}L^2B(B-1)^2\nu^2\\
&\quad+\left(4e^{1/5}\eta^3\overline{K}L^3B^3+12\eta^3\overline{K}^3B^3L^3\exp(e^{1/10}/5)e^{1/5}+\frac{\eta^3\overline{K}B^3L^3\exp(e^{1/10}/5)e^{2/5}}{50}\right)\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3\exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2\exp(e^{1/10}/5)e^{1/5}\\
&\leq\frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2+\frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2}+\frac{3\eta\overline{K}B}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3\exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2\exp(e^{1/10}/5)e^{1/5},
\end{aligned}\tag{97}$$

which completes the proof. □