# Federated Learning with Compression: Unified Analysis and Sharp Guarantees

Farzin Haddadpour[†] Mohammad Mahdi Kamani[‡] Aryan Mokhtari[§] Mehrdad Mahdavi[†]

[†]School of Electrical Engineering and Computer Science
[‡]College of Information Sciences and Technology
The Pennsylvania State University
{fxh18, mqk5591, mzm616}@psu.edu

[§]Department of Electrical and Computer Engineering
The University of Texas at Austin
mokhtari@austin.utexas.edu

**Abstract**

In federated learning, communication cost is often a critical bottleneck to scale up distributed optimization algorithms to collaboratively learn a model from millions of devices with potentially unreliable or limited communication and heterogeneous data distributions. Two notable trends to deal with the communication overhead of federated algorithms are *gradient compression* and *local computation with periodic communication*. Despite many attempts, characterizing the relationship between these two approaches has proven elusive. We address this by proposing a set of algorithms with periodical compressed (quantized or sparsified) communication and analyze their convergence properties in both homogeneous and heterogeneous local data distributions settings. For the homogeneous setting, our analysis improves existing bounds by providing tighter convergence rates for both *strongly convex* and *non-convex* objective functions. To mitigate data heterogeneity, we introduce a *local gradient tracking* scheme and obtain sharp convergence rates that match the best-known communication complexities without compression for convex, strongly convex, and nonconvex settings. We complement our theoretical results and demonstrate the effectiveness of our proposed methods by several experiments on real-world datasets.

## 1 Introduction

The primary obstacle towards scaling distributed optimization algorithms is the significant communication cost both in terms of the number of communication rounds and the amount of exchanged data per round. To significantly reduce the number of communication rounds, a practical solution is to trade-off local computation for less communication via periodic averaging [44, 57]. In particular, the local SGD algorithm [44, 49, 55] alternates between a fixed number of local updates and one step of synchronization which is shown to enjoy the same convergence rate as its fully synchronous counterpart, while significantly reducing the number of communication rounds.

A fundamentally different solution to scale up distributed optimization algorithms is to reduce the size of the communicated message per communication round. This problem is especially exacerbated in edge computing where the worker devices (e.g., smartphones or IoT devices) are remotely connected, and communication bandwidth and power resources are limited. For instance, ResNet [13] has more than 25 million parameters, so the communication cost of sending local models through a computer network could be prohibitive. The current methodology towards reducing the size of messages is to communicate compressed local gradients or models to the central server by utilizing a quantization operator [2, 5, 37, 47, 48, 52, 54], sparsification schema [3, 29, 45, 54], or composition of both [4].

Despite significant progress in improving both aspects of communication efficiency [5, 19, 45, 46], there still exists a huge gap in our understanding of these approaches in federated learning, in particular for the

| Reference | Objective function: Nonconvex | PL/Strongly Convex | General Convex | Unbounded gradient |
|---|---|---|---|---|
| [4] | $R=O\left(\frac{q+1}{\epsilon^{3/2}}\right)$ <br> $\tau=O\left(\frac{1}{m(q+1)\sqrt{\epsilon}}\right)$ | $R=O\left(\frac{q+1}{\sqrt{\epsilon}}\right)$ <br> $\tau=O\left(\frac{1}{m(q+1)\sqrt{\epsilon}}\right)$ | – | ✗ |
| [37] | $R=O\left(\frac{1}{\epsilon}\right)$ <br> $\tau=O\left(\frac{(\frac{q}{m})^2+1}{\epsilon}\right)$ | $R=O\left(m+\frac{q+1}{m\epsilon}\right)$ <br> $\tau=O\left(1\right)$ | – | ✔ |
| **Theorem 5.1** | $\boldsymbol{R=O\left(\frac{1}{\epsilon}\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\kappa\left(\frac{q}{m}+1\right)\log\left(\frac{1}{\epsilon}\right)\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\frac{1+\frac{q}{m}}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon^2}\right)}$ | ✔ |

Table 1: Comparison of results with compression and periodic averaging in the homogeneous setting. Here, $m$ is the number of devices, $q$ is compression distortion constant, $\kappa$ is condition number, $\epsilon$ is target accuracy, $R$ is the number of communication rounds, and $\tau$ is the number of local updates.

cases that both compression and periodic averaging techniques are applied simultaneously. In terms of reducing communication rounds, a few recent attempts were able to reduce the frequency of synchronizing locally evolving models [10, 20], which are not improvable in general [58]. This necessitates that further improvement in communication efficiency needs to be explored by reducing the size of communicated messages. We highlight that compressed communication is of further importance to accelerate training non-convex objectives as it requires significantly more communication rounds to converge compared to distributed convex optimization. Furthermore, most existing methods are analyzed for homogeneous data and our understanding of the efficiency of these methods in the heterogeneous case is lacking.

In light of the above issues, the key contribution of this paper is the introduction and analysis of simple variants of *local SGD*[1] *with compressed communication* without compromising the attainable guarantees. The proposed algorithmic ideas accommodate both homogeneous and heterogeneous data distributions settings with the obtained rates summarized in Table 1 and Table 2, respectively. In the homogeneous case, with a tight analysis of a simple quantized variant of local SGD, we show that not only our proposed method improves the complexity bounds for algorithms with compression (Table 1), but also outperforms the complexity bounds for non-compressed counterparts in terms of the number of communication rounds (Table 3). In the heterogeneous case, we argue that in the presence of compression (quantization or sparsification), locally updating models via local gradient information could lead to a significant drift among local models, which shed light on designing a quantized variant of local SGD that tracks local gradient information at local devices. We show that this simple gradient tracking idea leads to a method that outperforms state-of-the-art methods with compression for the heterogeneous setting (Table 2) and it can even compensate for the noise introduced by compression and lead to the best-known convergence rates for convex and non-convex settings under perfect communication, i.e., no compression (Table 4).

**Contributions.** We summarize the main contributions of this paper below:

- *Homogeneous local distributions:* To keep the analysis simple yet insightful, we start with a quantized variant of federated averaging algorithm and analyze its convergence for non-convex, strongly convex and general convex objectives. As demonstrated in Table 1, the obtained rates is novel for convex objectives to the best of our knowledge, and improves the best known bounds in [37] and [4] for general non-convex and strongly convex objectives, respectively.

- *Heterogeneous local distributions:* For the heterogeneous setting, we propose federated averaging with

[1]Based on the literature, noting the algorithmic similarity of Federated Averaging [33] and Local SGD [44], the main differences between them are the participation of clients and heterogeneity of local data distributions. In local SGD it is usually assumed that all of the clients are involved in communication, whereas in federated averaging a randomly selected subset of clients participates at averaging. Also, federated averaging is commonly used to reflect the data heterogeneity, which is a key ingredient in our analysis as well. For simplicity, we do not differentiate between these two terms and use them interchangeably.

| Reference | Objective function | | | Unbounded gradient |
|---|---|---|---|---|
| | Nonconvex | PL/Strongly Convex | General Convex | |
| [37] | $R=O\left(\frac{q+1}{\epsilon^{3/2}}\right)$ <br> $\tau=O\left(\frac{1}{m(q+1)\sqrt{\epsilon}}\right)$ | $R=O\left(\frac{q+1}{\sqrt{\epsilon}}\right)$ <br> $\tau=O\left(\frac{1}{m(q+1)\sqrt{\epsilon}}\right)$ | – | ✗ |
| **Theorem 5.2** | $\boldsymbol{R=O\left(\frac{q+1}{\epsilon}\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\kappa\,(q+1)\log\left(\frac{1}{\epsilon}\right)\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\frac{1+q}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)}$ <br> $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon^2}\right)}$ | ✔ |

Table 2: Comparison of results with compression and periodic averaging in the heterogeneous setting.

compression and local gradient tracking, dubbed as `FedCOMGATE` algorithm, and establish its convergence rates for general non-convex, strongly convex or PL, and convex objectives. The obtained rates improve upon the results reported in [4] for general non-convex and strongly-convex objectives. The obtained rates for general convex functions are novel to the best of our knowledge.

- We verify our theoretical results through various extensive experiments on different real federated datasets that demonstrate the practical efficacy of our methods.

# 2 Problem Setup

In this paper we focus on a federated architecture, where $m$ users aim to learn a global model in a collaborative manner without exchanging their data points with each other. Moreover, we assume that users (computing units) can only exchange information via a central unit (server) which is connected to all users. The optimization problem that the users try to solve can be written as

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) \triangleq \frac{1}{m}\sum_{j=1}^{m} f_j(\boldsymbol{w}) \tag{1}$$

where $f_j:\mathbb{R}^d\to\mathbb{R}$ is loss function corresponding to user $j$. We further assume that the local objective function of each user $j$ is the loss over the set of data points of node $j$, i.e.,

$$f_j(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}_j}[\ell_j(\boldsymbol{w},\boldsymbol{z})], \tag{2}$$

where $\boldsymbol{z}$ is a random variable with probability distribution $\mathcal{P}_j$ and the loss function $\ell_j$ measures how well the model performs. Here $\mathcal{P}_j$ can be considered as the underlying distribution of node $j$ for generating data points, and realizations of the random variable $\boldsymbol{z}$ are the data points of node $j$. For instance, in a supervised learning case each element sample point $\boldsymbol{z}_i$ corresponds to a pair of input vector (feature vector) $\boldsymbol{x}_i$ and its label $y_i$. In this case, $\ell_j(\boldsymbol{w},\boldsymbol{z}_i)=\ell_j(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$ measures how well the model $\boldsymbol{w}$ performs in predicting the label of $\boldsymbol{x}_i$ which is $y_i$. Note that the probability distributions of users may not be necessarily identical. In fact, through the paper, we study two settings (i) homogeneous setting in which all the probability distributions and loss functions are identical, i.e., $(\mathcal{P}_1=\ldots=\mathcal{P}_m)$ and $(\ell_1=\ldots=\ell_m)$; and (ii) heterogeneous setting in which the users' distributions and loss functions could be different.

# 3 Federated Averaging with Compression[2]

In this section, we propose a generalized version of the local stochastic gradient descent (SGD) method for federated learning which uses compressed signals to reduce the overall communication overhead of solving problem (1). The proposed federated averaging with compression (`FedCOM`) is designed for homogeneous settings where the probability distributions and loss functions of the users are identical. `FedCOM` differs from

[2] Generalized Compressed Local SGD

**Algorithm 1:** `FedCOM`$(R, \tau, \eta, \gamma)$: Federated Averaging with Compression and Decoupled Rates

**Inputs:** Number of communication rounds $R$, number of local updates $\tau$, learning rates $\gamma$ and $\eta$, initial global model $\boldsymbol{w}^{(0)}$

**for** $r = 0, \ldots, R-1$ **do**

  **for** *each client* $j \in [m]$ **do in parallel**

    Set $\boldsymbol{w}_j^{(0,r)} = \boldsymbol{w}^{(r)}$

    **for** $c = 0, \ldots, \tau - 1$ **do**

      Sample a minibatch $\mathcal{Z}_j^{(c,r)}$ and compute $\tilde{\mathbf{g}}_j^{(c,r)} \triangleq \nabla f_j(\boldsymbol{w}_j^{(c,r)}; \mathcal{Z}_j^{(c,r)})$

      $\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\, \tilde{\mathbf{g}}_j^{(c,r)}$

    **end**

    Device **sends** $\boldsymbol{\Delta}_{j,q}^{(r)} = Q((\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)})/\eta)$ back to the server

  **end**

  Server **computes** $\boldsymbol{\Delta}_q^{(r)} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{\Delta}_{j,q}^{(r)}$

  Server **computes** $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma\boldsymbol{\Delta}_q^{(r)}$ and **broadcasts** to all devices

**end**

standard local SGD methods [44, 49, 55] in two major aspects. First, it uses compressed messages for uplink communication. Second, at the central node, the new global model is a convex combination of the previous global model and the average of updated local models of users. We show that `FedCOM` converges faster than state-of-the-art methods in a homogeneous setting by periodic averaging, local and global learning rates, and compressed communications.

To formally present the steps of `FedCOM`, consider $R$ as the rounds of communication between server and users, and $\tau$ as the number of local updates performed between two consecutive communication rounds. Further, define $\boldsymbol{w}^{(r)}$ as the model at the master at the $r$-th round of communication. At each round $r$, the server sends the global model $\boldsymbol{w}^{(r)}$ to the users (clients). Then, each user $j$ computes its local stochastic gradient and updates the model by following the update of SGD for $\tau$ iterations. Specifically, at communication round $r$, user $j$ follows the update

$$\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\, \tilde{\mathbf{g}}_j^{(c,r)}, \qquad \text{for} \quad c = 0, \ldots, \tau - 1. \tag{3}$$

Here, $\boldsymbol{w}_j^{(c,r)}$ is the model at node $j$ and round $r$ after $c$ local updates, $\tilde{\mathbf{g}}_j^{(c,r)} := \nabla f_j(\boldsymbol{w}_j^{(c,r)}; \mathcal{Z}_j^{(c,r)}) := \frac{1}{b_j}\sum_{\mathbf{z}\in\mathcal{Z}_j^{(c,r)}} \nabla \ell_j(\boldsymbol{w}_j^{(c,r)}, \mathbf{z})$ is a stochastic gradient of $f_j$ evaluated using the mini-batch $\mathcal{Z}_j^{(c,r)} := \{\boldsymbol{z}_{j,1}^{(c,r)}, \ldots, \boldsymbol{z}_{j,b_j}^{(c,r)}\}$ of size $b_j$, and $\eta$ is the learning rate. The output of this $\tau$ recursive updates for node $j$ at round $r$ is $\boldsymbol{w}_j^{(\tau,r)}$. After computing the local models, each user $j$ sends a compressed version of $(\boldsymbol{w}_j^{(\tau,r)} - \boldsymbol{w}_j^{(r)})/\eta$ to the central node by applying a compression operator $Q(\cdot)$. Note that the compressed signal $\boldsymbol{\Delta}_{j,q}^{(r)} \triangleq Q((\boldsymbol{w}_j^{(\tau,r)} - \boldsymbol{w}_j^{(r)})/\eta)$ indicates a normalized version of the difference between the input and output of the local SGD process at round $r$ at node $j$, which is equal to the aggregation of all local SGD directions, i.e., $(\boldsymbol{w}_j^{(\tau,r)} - \boldsymbol{w}_j^{(r)})/\eta = \sum_{c=0}^{\tau} \tilde{\mathbf{g}}_j^{(c,r)}$. Once, the server receives the compressed signals $\{\boldsymbol{\Delta}_{j,q}^{(r)}\}_{j=1}^{j=m}$, it computes the new global model according to the update

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \frac{\eta\gamma}{m}\sum_{j=1}^{m} \boldsymbol{\Delta}_{j,q}^{(r)}, \tag{4}$$

where $\gamma$ is the global learning rate. The steps of the proposed `FedCOM` algorithm are summarized in Algorithm 1.

**Remark 1.** *Note that by setting $\gamma = 1$ in (4), `FedCOM` boils down to the FedPAQ algorithm proposed in [37], and if we further remove the compression scheme then we recover FedAvg [49]. Note that in both FedAvg and its vanilla quantized variant FedPAQ, the new global model is the average of local models (if we ignore the error of compression for FedPAQ), while in `FedCOM` the new global model is a linear combination of the* *previous global model and the average of updated local models, due to the extra parameter $\gamma$. We show that by adding this modification and properly choosing $\gamma$, `FedCOM` improves the complexity bounds of FedPAQ for both strongly convex and non-convex settings. Note that the update in (4) can also be interpreted as running a global SGD update on master's model by descending towards the average of aggregated local gradient directions with stepsize $\eta\gamma$. Specifically, if we assume perfect communication (ignoring the quantization) then we obtain that the new global model is given by $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau}\tilde{\mathbf{g}}_j^{(c,r)}$.*

# 4 Compressed Local SGD with Local Gradient Tracking

In the previous section, we introduced a relatively simple algorithm called `FedCOM` for homogeneous settings, where the probability distributions of the users are identical. Although `FedCOM` both theoretically (Section 5) and numerically (Section 6) performs well for homogeneous settings, its performance is not satisfactory in heterogeneous settings where the probability distributions of users are different. This is due to the fact that the updates of `FedCOM` heavily depend on the local SGD directions. In a homogeneous setting, following local gradient directions leads to a good global model as all samples are drawn from the same distribution and the local gradient direction is a good estimate of the global function gradient. However, in a heterogeneous setting, updating local models only based on local gradient information could lead to an arbitrary poor performance as the local gradient directions could be very different from the global gradient direction.

To address this issue, in this section we propose a novel variant of federated averaging with compression and local gradient tracking (`FedCOMGATE`) for heterogeneous settings. The main difference between `FedCOM` and `FedCOMGATE` is the idea of local gradient tracking that ensures that each node uses an estimate of the global gradient direction to locally update its model. To estimate global gradient direction nodes also require access to the average of local models which means that in `FedCOMGATE` in addition to sending the global updates master also needs to broadcast the average of $\boldsymbol{\Delta}_{j,q}^{(r)} \triangleq Q((\boldsymbol{w}_j^{(\tau,r)} - \boldsymbol{w}_j^{(r)})/\eta)$, shown by $\boldsymbol{\Delta}_q^{(r)} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\Delta}_{j,q}^{(r)}$ to devices.

To present `FedCOMGATE`, consider $\delta_j$ as a sequence at node $j$ that is designed to track the difference between the local gradient direction and the global gradient direction (the direction obtained by incorporating gradient information of all users). At round $r$, each worker $j$ updates its local sequence $\delta_j$ based on the update

$$\delta_j^{(r+1)} = \delta_j^{(r)} + \frac{1}{\tau}\left(\boldsymbol{\Delta}_{j,q}^{(r)} - \boldsymbol{\Delta}_q^{(r)}\right), \tag{5}$$

where $\boldsymbol{\Delta}_{j,q}^{(r)}$ is the quantized version of the accumulation of the gradients at node $j$ from the previous round and $\boldsymbol{\Delta}_q^{(r)}$ is the average of $\boldsymbol{\Delta}_{j,q}^{(r)}$. Once the correction vector $\delta_j^{(r)}$ is computed, each node $j$ runs a corrected local update for $\tau$ rounds based on the update

$$\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\,\tilde{\boldsymbol{d}}_{j,q}^{(c,r)} = \boldsymbol{w}_j^{(c,r)} - \eta(\tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}), \qquad \text{for} \quad c = 0, \ldots, \tau-1, \tag{6}$$

where $\tilde{\mathbf{g}}_j^{(c,r)} \triangleq \nabla f_j(\boldsymbol{w}_j^{(c,r)}\mathcal{Z}_j^{(c,r)})$ is the stochastic gradient of node $j$ at round $r$ for the $c$-th local update. In the above update the local descent direction $\tilde{\boldsymbol{d}}_{j,q}^{(c,r)}$ is defined as the difference the local stochastic gradient $\tilde{\mathbf{g}}_j^{(c,r)}$ and the correction vector $\delta_j^{(r)}$ which aims to track the difference between local and global gradient directions. Note that for all $\tau$ local updates at round $r$, the vector $\delta_j^{(r)}$ is fixed while the local stochastic gradient $\tilde{\mathbf{g}}_j^{(c,r)}$ is computed via fresh samples for each local update.

Once the local models $\boldsymbol{w}_j^{(\tau,r)}$ are computed, nodes send their quantized accumulation of gradients $\boldsymbol{\Delta}_{j,q}^{(r)}$ to the server. Then, the server uses this information to compute the average update $\boldsymbol{\Delta}_q^{(r)} \triangleq \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\Delta}_{j,q}^{(r)}$ and broadcasts it to the devices. Moreover, the server utilizes $\boldsymbol{\Delta}_q^{(r)}$ to compute the new global model $\boldsymbol{w}^{(r+1)}$ according to (4). The steps of `FedCOMGATE` are outlined in Algorithm 2.

**Algorithm 2:** `FedCOMGATE`$(R, \tau, \eta, \gamma)$: Federated Averaging with Compression and Gradient Tracking

**Inputs:** Number of communication rounds $R$, number of local updates $\tau$, learning rates $\gamma$ and $\eta$, initial global model $\boldsymbol{w}^{(0)}$, initial gradient tracking $\delta_j^{(0)} = \mathbf{0},\ \forall j \in [m]$

**for** $r = 0, \ldots, R-1$ **do**
  **for** *each client* $j \in [m]$ **do in parallel**
    Set $\boldsymbol{w}_j^{(0,r)} = \boldsymbol{w}^{(r)}$
    **for** $c = 0, \ldots, \tau - 1$ **do**
      Set $\tilde{\boldsymbol{d}}_{j,q}^{(c,r)} = \tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}$ where $\tilde{\mathbf{g}}_j^{(c,r)} \triangleq \nabla f_j(\boldsymbol{w}_j^{(c,r)}; \mathcal{Z}_j^{(c,r)})$
      $\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\, \tilde{\boldsymbol{d}}_{j,q}^{(c,r)}$
    **end**
    Device **sends** $\boldsymbol{\Delta}_{j,q}^{(r)} = Q((\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)})/\eta)$ to the server
    Device **updates** $\delta_j^{(r+1)} = \delta_j^{(r)} + \frac{1}{\tau}(\boldsymbol{\Delta}_{j,q}^{(r)} - \boldsymbol{\Delta}_q^{(r)})$ `// Gradient Tracking`
  **end**
  Server **computes** $\boldsymbol{\Delta}_q^{(r)} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\Delta}_{j,q}^{(r)}$ and **broadcasts** back to all devices
  Server **computes** $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma\boldsymbol{\Delta}_q^{(r)}$ and **broadcasts** to all devices
**end**

**Comparison with SCAFFOLD [19] and VRL-SGD in [26].** From an algorithmic standpoint, in comparison to the SCAFFOLD method proposed in [19] , in addition to the fact that we use compressed signals to further reduce the communication overhead, we would like to highlight that our algorithm is much simpler and does not require any extra control variable (see Eq. (4) and Eq. (5) in [19] for more details). Also, since we do not use an extra control variable, the extension of our convergence analysis to the case where a subset of devices participate at each communication round is straightforward and for clarity, we do not include analysis with device sampling. Yet, we shall study the impact of device sampling empirically (see Figure 5 in Section 6 and Algorithm 4 in Appendix B). In comparison to [26] which employs an explicit variance reduction component, if we let $Q(\boldsymbol{x}) = \boldsymbol{x}$ (case of no quantization), our algorithm reduces to a generalization of algorithm in [26] with distinct local and global learning rates. We note that for the case of $\gamma = 1$ and $Q(\boldsymbol{x}) = \boldsymbol{x}$ the `FedCOMGATE`$(\tau, \eta, \gamma = 1)$ reduces to the federated algorithm proposed in [26] with minor distinction that our algorithm's output is the global model at the server.

# 5 Convergence Analysis

Next, we present the convergence analysis of our proposed methods. First, we state our assumptions.

**Assumption 1** (Smoothness and Lower Boundedness)**.** *The local objective function $f_j(\cdot)$ of $j$th device is differentiable for $j \in [m]$ and $L$-smooth, i.e., $\|\nabla f_j(\boldsymbol{u}) - \nabla f_j(\boldsymbol{v})\| \le L\|\boldsymbol{u} - \boldsymbol{v}\|,\ \forall\ \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$. Moreover, the optimal value of objective function $f(\cdot)$ is bounded below by $f^* = \min_{\boldsymbol{w}} f(\boldsymbol{w}) > -\infty$.*

**Assumption 2.** *The output of the compression operator $Q(\boldsymbol{x})$ is an unbisased estimator of its input $\boldsymbol{x}$, and and its variance grows with the squared of the squared of $\ell_2$-norm of its argument, i.e., $\mathbb{E}[Q(\boldsymbol{x})|\boldsymbol{x}] = \boldsymbol{x}$ and $\mathbb{E}[\|Q(\boldsymbol{x}) - \boldsymbol{x}\|^2|\boldsymbol{x}] \le q\,\|\boldsymbol{x}\|^2$.*

The conditions in Assumptions 1-2 are customary in the analysis of compressed federated learning methods, and they will all be assumed in all of our results. We should also add that several quantization approaches and sparsification techniques satisfy the condition in Assumption 2. For examples of such compression schemes we refer the reader to [4, 14]. We report our results for three different class of loss functions: (i) nonconvex (ii) convex (iii) non-convex Polyak-Łojasiewicz (PL). Indeed, as any $\mu$-strongly convex is $\mu$-PL [18], our results for the PL case automatically hold for strongly convex functions.

### 5.1 Convergence of `FedCOM` in the homogeneous data distribution setting

Now we focus on the homogeneous case in which the stochastic local gradient of each worker is an unbiased estimator of the global gradient.

**Assumption 3** (Bounded Variance)**.** *For all $j \in [m]$, we can sample an independent mini-batch $\mathcal{Z}_j$ of size $|\mathcal{Z}_j^{(c,r)}| = b$ and compute an unbiased stochastic gradient $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{w}; \mathcal{Z}_j), \mathbb{E}_{\mathcal{Z}_j}[\tilde{\mathbf{g}}_j] = \nabla f(\boldsymbol{w}) = \mathbf{g}$. Moreover, their variance is bounded above by a constant $\sigma^2$, i.e., $\mathbb{E}_{\mathcal{Z}_j}\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}\|^2\right] \leq \sigma^2$.*

In the following theorem, we state our main theoretical results for the `FedCOM` algorithm in the homogeneous setting.

**Theorem 5.1.** *Consider `FedCOM` in Algorithm 1. Suppose that the conditions in Assumptions 1-3 hold. If the local data distributions of all users are identical (homogeneous setting), then we have*

- ***Nonconvex:*** *By choosing stepsizes as $\eta = \frac{1}{L\gamma}\sqrt{\frac{m}{R\tau\left(\frac{q}{m}+1\right)}}$ and $\gamma \geq m$, the sequence of iterates satisfies $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ if we set $R = O\left(\frac{1}{\epsilon}\right)$ and $\tau = O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)$.*
- ***Strongly convex or PL:*** *By choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{q}{m}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq \epsilon$ if we set $R = O\left(\left(\frac{q}{m}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon}\right)$.*
- ***Convex:*** *By choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{q}{p}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq \epsilon$ if we set $R = O\left(\frac{L\left(1+\frac{q}{m}\right)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon^2}\right)$.*

Theorem 5.1 characterizes the number of required local updates $\tau$ and communication rounds $R$ to achieve an $\epsilon$-first-order stationary point for the nonconvex setting and an $\epsilon$-suboptimal solution for convex and strongly convex settings, when we are in a homogeneous case. A few important observations follow. First, in all three results the dependency of $\tau$ and $R$ on the variance of compression scheme $q$ is scaled down by a factor of $1/m$. Hence, by cooperative learning the users are able to lower the effect of the noise induced by the compression scheme. Second, in all three cases, the number of local updates $\tau$ required for achieving a specific accuracy is proportional to $1/m$. In the homogeneous setting, this result is expected since we have $m$ machines and the number of samples used per local update is $m$ times of the case that only a single machine runs local SGD. As a result, the overall number of required local updates scales inversely by the number of machines $m$. Third, in all three cases, the dependency of communication rounds $R$ on the required accuracy $\epsilon$ matches the number of required updates for solving that problem in centralized deterministic settings. For instance, in a centralized nonconvex setting, to achieve a point that satisfies $\|\nabla f(\boldsymbol{w})\|^2 \leq \epsilon$ we need $O(1/\epsilon)$ gradient updates for deterministic case and $O(1/\epsilon^2)$ SGD updates for the stochastic case. It is interesting that running $\tau = O(1/\epsilon)$ local updates controls the noise of stochastic gradients and the number of communication rounds $R = O(1/\epsilon)$ stays same as the centralized deterministic case. Similar observations hold for convex (upto a log factor) and strongly convex cases.

**Remark 2.** *While in the bound obtained in Theorem 5.1 for general non-convex objectives indicates that achieving a convergence rate of $\epsilon$ requires $R = O\left(\frac{\frac{q}{m}+1}{\epsilon}\right)$ communication rounds with $\tau = O\left(\frac{1}{m\epsilon}\right)$ local updates, in Remark 6 in Appendix D, we show that the same rate can be achieved with $R = O\left(\frac{1}{\epsilon}\right)$ and $\tau = O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)$. We note that this is an interesting observation demonstrating that the noise of quantiziation can be compensated with higher number of local steps $\tau$.*

**Remark 3.** *The results for `FedCOM` improve the complexity bounds for other federated leaning methods with compression (in the homogeneous setting) that are proposed in [37] and [4]. Check Table 1 for more details.*

| Reference | Objective function | | |
|---|---|---|---|
| | Nonconvex | PL/Strongly Convex | General Convex |
| [10] | – | $R = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{1}{3}}\right)$ $\tau = O\left(\frac{1}{m\epsilon}\right)$ | – |
| [20] | – | $R = O\left(m\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{m^2\epsilon}\right)$ | $R = O\left(\frac{m}{\epsilon}\right)$ $\tau = O\left(\frac{1}{m^2\epsilon}\right)$ |
| [49] | $R = O\left(\frac{m}{\epsilon}\right)$ $\tau = O\left(\frac{1}{m^2\epsilon}\right)$ | – | – |
| **Theorem 5.1** | $\boldsymbol{R = O\left(\frac{1}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R = O\left(\kappa \log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R = O\left(\frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon^2}\right)}$ |

Table 3: Comparison of `FedCOM` with results that use periodic averaging but do not utilize compression, i.e., $q = 0$, in the homogeneous setting.

**Remark 4.** *To show the tightness of our result for* `FedCOM`*, we also compare its results with schemes without compression, $q = 0$, developed for homogeneous settings, shown in Table 3. As we observe, the number of required communication rounds for* `FedCOM` *without compression in convex, strongly convex, and nonconvex settings are smaller than the best-known rates for each setting by a factor of $\frac{1}{m}$.*

## 5.2 Convergence of `FedCOMGATE` in the data heterogeneous setting

Next, we report our results for `FedCOMGATE` in the heterogeneous setting. We consider a less strict assumption compared to Assumption 3, that the stochastic gradient of each user is an unbiased estimator of its local gradient with bounded variance.

**Assumption 4** (Bounded Local Variance)**.** *For all $j \in [m]$, we can sample an independent mini-batch $\mathcal{Z}_j$ of size $|\mathcal{Z}_j| = b$ and compute an unbiased stochastic gradient $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{w}; \mathcal{Z}_j), \mathbb{E}_\xi[\tilde{\mathbf{g}}_j] = \nabla f_j(\boldsymbol{w}) = \mathbf{g}_j$. Moreover, the variance of local stochastic gradients is bounded above by a constant $\sigma^2$, i.e., $\mathbb{E}_\xi\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}_j\|^2\right] \leq \sigma^2$.*

**Assumption 5.** *The compression scheme $Q$ for the heterogeneous data distribution setting satisfies the following condition $\mathbb{E}_Q[\|\frac{1}{m}\sum_{j=1}^m Q(\boldsymbol{x}_j)\|^2 - \|Q(\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_j)\|^2] \leq G_q$.*

Note that the condition in Assumption 4 is very general and only ensures that the local stochastic gradients are unbiased estimators of local gradients and have bounded variance.

For the case of no compression, $Q(\boldsymbol{x}) = \boldsymbol{x}$, the compression error in Assumption 5 becomes naturally $G_q = 0$. We highlight that this assumption is only needed in the heterogeneous setting, and since both of the terms in the argument of expectation depend on the quantization, this assumption can be seen as a weaker version of the gradient diversity assumptions in the convergence analysis of heterogeneous settings. To show how this assumption holds in practice, we run an experiment on the MNIST dataset using `FedCOMGATE` algorithm with quantizing gradients from 32 bits floating-point to 8 bits integer. In Figure 1 we plot changes in $G_q$ quantity through this experiment. It shows that $G_q$ is decreasing as we proceed with the training, simply because the $\ell_2$-norm of the updated vector is going to zero. Note that the quantity of $G_q$ could be even negative as illustrated in Figure 1. For more details on the experiment see Section 6.

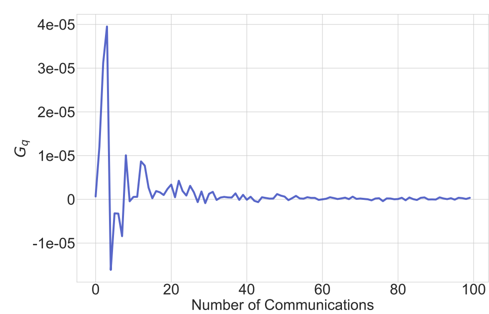



Figure 1: The error of quantization measured by Assumption 5 for `FedCOMGATE` with the MNIST dataset applied to a MLP model. We quantized updates $\boldsymbol{\Delta}_j$s from 32 bits floating-point to 8 bits integer.

Next, we present our main theoretical results for the `FedCOMGATE` method in the heterogeneous setting.

| Reference | Objective function: Nonconvex | PL/Strongly Convex | General Convex | Extra control variable |
|---|---|---|---|---|
| [19] | $R=O\left(\frac{1}{\epsilon}\right)$ $\tau=O\left(\frac{1}{m\epsilon}\right)$ | $R=O\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau=O\left(\frac{1}{m\epsilon}\right)$ | $R=O\left(\frac{1}{\epsilon}\right)$ $\tau=O\left(\frac{1}{m\epsilon}\right)$ | Yes |
| [20] | − | − | $R=O\left(\frac{1}{\epsilon^{1.5}}\right)$ $\tau=O\left(\frac{1}{m\epsilon^{0.5}}\right)$ | No |
| [26] | $R=O\left(\frac{m}{\epsilon}\right)$ $\tau=O\left(\frac{1}{m^2\epsilon}\right)$ | − | − | No |
| **Theorem 5.2** | $\boldsymbol{R=O\left(\frac{1}{\epsilon}\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon}\right)}$ | $\boldsymbol{R=O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon^2}\right)}$ | **No** |

Table 4: Comparison of `FedCOMGATE` with results that use periodic averaging but do not utilize compression, i.e., $q=0$, in the heterogeneous setting.

**Theorem 5.2.** *Consider `FedCOMGATE` in Algorithm 2. If Assumptions 1, 2, 4 and 5 hold, then even for the case the local data distribution of users are different (heterogeneous setting) we have*

- ***Non-convex:*** *By choosing stepsizes as* $\eta=\frac{1}{L\gamma}\sqrt{\frac{m}{R\tau(q+1)}}$ *and* $\gamma\geq m$, *we obtain that the iterates satsify* $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2\leq\epsilon$ *if we set* $R=O\left(\frac{q+1}{\epsilon}\right)$ *and* $\tau=O\left(\frac{1}{m\epsilon}\right)$.
- ***Strongly convex or PL:*** *By choosing stepsizes as* $\eta=\frac{1}{2L\left(\frac{q}{m}+1\right)\tau\gamma}$ *and* $\gamma\geq\sqrt{m\tau}$, *we obtain that the iterates satisfy* $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big]\leq\epsilon$ *if we set* $R=O\left((q+1)\,\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau=O\left(\frac{1}{m\epsilon}\right)$.
- ***Convex:*** *By choosing stepsizes as* $\eta=\frac{1}{2L(q+1)\tau\gamma}$ *and* $\gamma\geq\sqrt{m\tau}$, *we obtain that the iterates satisfy* $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big]\leq\epsilon$ *if we set* $R=O\left(\frac{L(1+q)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau=O\left(\frac{1}{m\epsilon^2}\right)$.

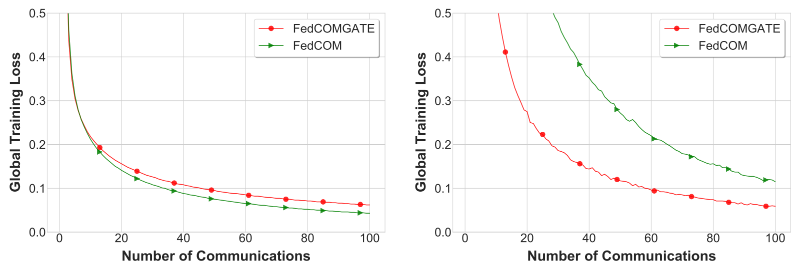
The implications of Theorem 5.2 are similar to the ones for Theorem 5.1. Yet, unlike the homogeneous setting, the compression variance $q$ does not scale down by a factor of $1/m$. We emphasize that similar to the homogeneous case, in all three cases, the dependency of $R$ on $\epsilon$ matches the number of required update for solving the problem in centralized fashion.

**Remark 5.** *To show the tightness of our result for `FedCOMGATE`, we also compare its complexity bounds with other schemes without compression,* $q=0$, *for heterogeneous settings, summarized in Table 4. As we observe for convex, strongly convex and nonconvex settings, the results for `FedCOMGATE` without compression (`FedGATE`) match the best-known complexity bounds for these settings (upto a log factor).*

# 6 Experiments

In this section, we empirically validate the performance of proposed algorithms. We compare our methods with FedAvg [33], its quantized version, FedPAQ [37], and SCAFFOLD [19] for heterogeneous federated learning. In addition, we present a variant of our algorithm without compression dubbed as `FedGATE`. The details of `FedGATE` is described in Algorithm 3, which is deferred to Appendix B. Also, we present a variant of our algorithm with the client sampling in Algorithm 4, which is used in its corresponding experiment part.

**Setup.** We implement our algorithms on the `Distributed` library of `PyTorch` [35], using Message Passing Interface (`MPI`), in order to simulate the real-world collaborative learning scenarios such as the one in federated learning. We run the experiments on a HPC cluster with 3 `Intel Xeon E5-2695` CPUs, each of which with 28 processes. On an HPC cluster like this, the communication between nodes and processes are efficient, which is far away from the real-world scenarios, where the main bottleneck is the communication in the system.




(a) Homogeneous



(b) Heterogeneous

Figure 2: Comparing Algorithm 1 and Algorithm 2 for homogeneous and heterogeneous data distributions of the `MNIST` dataset. In heterogeneous distribution, `FedCOM` suffers from a residual error.

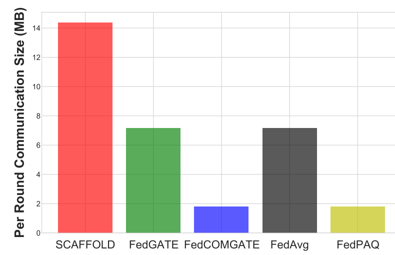



Figure 3: Communication cost at each round for the `CIFAR10` dataset with a 2-layer MLP.

Hence, we simulate the real-world communication channel delay, which is proportional to the size of the message. For this experiment we use three main datasets: `MNIST`[3], `CIFAR10`[4], and `Fashion MNIST`[5].

For each experiment, we have 100 devices communicating with the server. Each experiment runs for 100 rounds of communication between clients and the server, and we report the global model loss on the training data averaged over all clients. For `MNIST` and `Fashion MNIST` we use an MLP model with two hidden layers, each with 200 neurons with ReLU activations. For the `CIFAR10` dataset, we use the same MLP model, each layer with 500 neurons.

**Homogeneous data distribution.** The `FedCOM` algorithm is best suited for the homogeneous case, while in the heterogeneous setting it suffers from a residual error. That is why we use gradient tracking in the `FedCOMGATE` algorithm. This error can be seen in Figure 2 for `MNIST` data, where in Figure 2(a) data is distributed homogeneously among devices, and in Figure 2(b) each device has access to only 2 classes in the dataset. The results indicate that we need gradient tracking in `FedCOMGATE` to deal with heterogeneity.

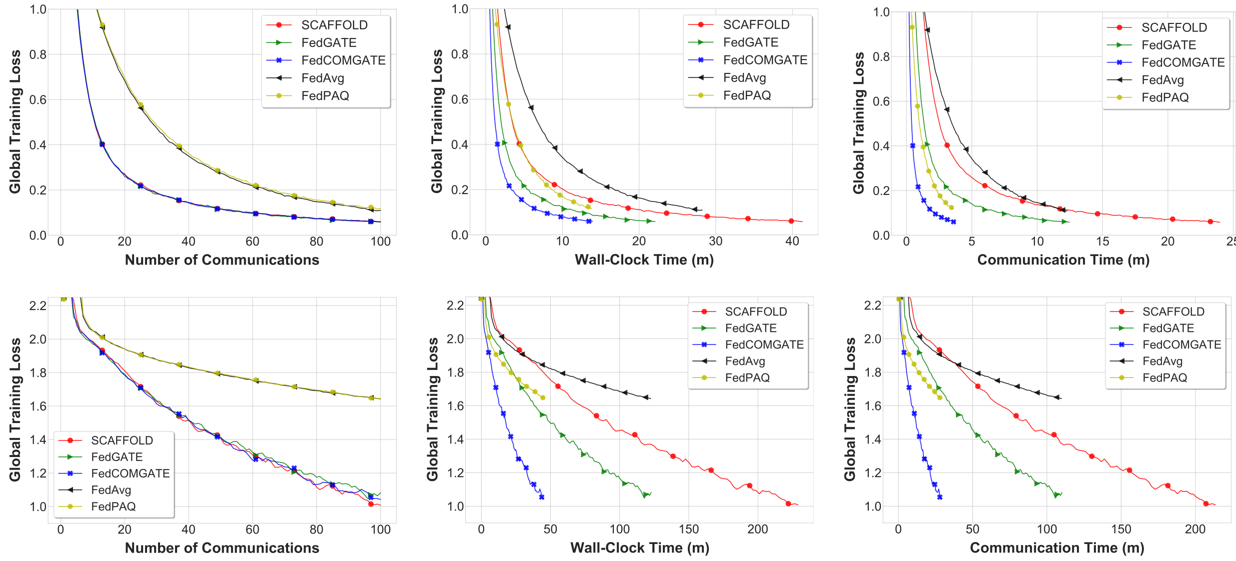
**Heterogeneous data distribution.** To generate heterogeneous data that resembles a real federated learning setup, we will follow a similar approach as in [33]. In this regard, we will distribute the data among clients in a way that each client only has data from two classes, which is highly heterogeneous. The idea behind `FedCOMGATE` is similar to the one in SCAFFOLD, except in `FedCOMGATE` we only have one control variable that gets updated using normal updates in FedAvg. In contrast, SCAFFOLD has two control variables and requires to update the global model and server control variable at each round. Hence, SCAFFOLD communicates at least twice the size as `FedCOMGATE` with the server at each round, when we do not use any compression. With compression, say $4\times$ quantization, we can substantially reduce the communication cost, say $8\times$, with respect to SCAFFOLD, while preserving the same convergence rate. To compare their communication cost, in Figure 3 we show the size of variables each device in each algorithm communicates with the server (for the gathering or uplink only, since the broadcasting or downlink time is negligible compared to the gathering) for the `CIFAR10` dataset with an MLP model that has 2 hidden layers, each with 500 neurons. For `FedCOMGATE` and FedPAQ we quantize the updates from 32 bits floating-point to 8 bits integer.

To show the effect of this communication size on the real-time convergence of each algorithm, we run each of them on the `MNIST` and the `CIFAR10` datasets with MLP models as described before. The data is distributed heterogeneously among clients, where each one has access to only 2 classes. Figure 4(a) shows the global model loss on training data on each communication round. FedAvg and FedPAQ are very close to each other on each communication round, whereas `FedCOMGATE`, its normal version without compression `FedGATE`, and SCAFFOLD are performing similarly based on communication rounds. Figure 4(b) shows this

[3] http://yann.lecun.com/exdb/mnist/

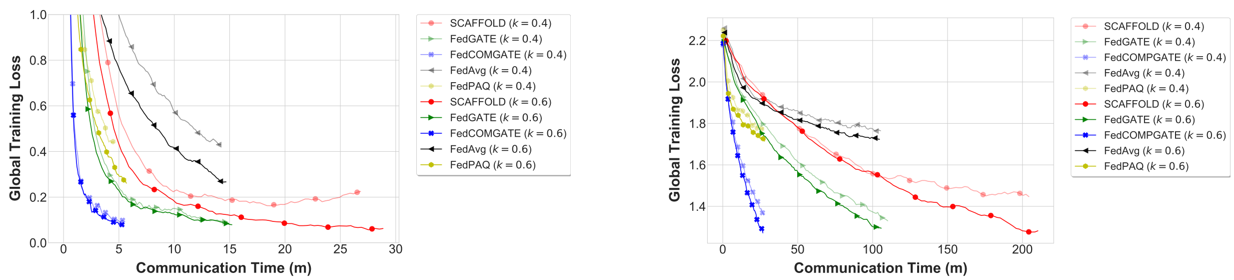
[4] https://www.cs.toronto.edu/~kriz/cifar.html

[5] https://github.com/zalandoresearch/fashion-mnist




(a) (b) (c)

Figure 4: Comparing FedCOMGATE and FedGATE with FedAvg [33], FedPAQ [37], and SCAFFOLD [19] on the MNIST (first row) and the CIFAR10 (second row) datasets. Both FedCOMGATE and FedGATE outperform other algorithms in terms of communication time and convergence rate.

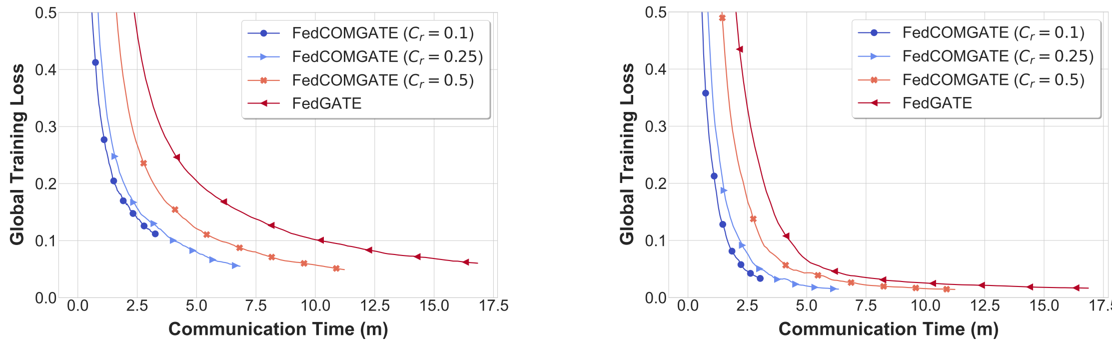



(a) MNIST Dataset (b) CIFAR10 Dataset

Figure 5: Comparing the effect of sampling on different algorithms. We use two datasets: the MNIST and the CIFAR10 datasets. We use an MLP with 2 layers for all the datasets, with 200 neurons per layer for the MNIST, and 500 neurons per layer for the CIFAR10. FedGATE and FedCOMGATE seem to be more robust against client sampling.

loss based on the wall-clock time (computation + communication) and Figure 4(c) shows this loss based on communication time only. Both Figures 4(b) and 4(c) clearly demonstrate the effectiveness of proposed algorithms. Especially, the FedCOMGATE algorithm superbly outperforms other algorithms where the model size is relatively large.

**Effect of sampling.** In this set of experiments, based on Algorithm 4, we assume that only $k \in (0, 1]$ portion of the devices could be sampled and exchange information with the server at each round. Indeed, a lower value of $k$ implies that fewer nodes are active at each round and therefore the communication overhead is lower. However, it could possibly lead to a slower convergence rate and extra communications rounds to achieve a specific accuracy. We formally study the effect of $k$ on the convergence of FedCOMGATE and its variant without compression FedGATE, and compare their performance with other federated methods with and without compression in Figure 5. As it can be inferred, when we decrease $k$ or the participation rate,




(a) MNIST Dataset

(b) Fashion MNIST Dataset

Figure 6: The effect of sparsification with memory on the `FedCOMGATE` algorithm used for the training of the `MNIST` and the `Fashion MNIST` datasets. We can achieve almost similar results as the algorithm without compression (`FedGATE`) with some compression rates. Decreasing the size of communication will speed up the training, in the cost of increasing a residual error as it is evident for the case with $C_r = 0.1$.

generally, the performance of the model degrades with the same number of communication rounds. However, the amount of degradation might vary among different algorithms. As it is depicted in Figure 5, the proposed `FedCOMGATE` algorithm and its version without compression, `FedGATE`, are quite robust against decreasing the participation rate between clients with respect to other algorithms such as SCAFFOLD and FedPAQ.

**Compression via sparsification.** Another approach to compress the gradient updates is sparsification. This method has been vastly used in distributed training of machine learning models [1,45,51]. Using a simple sparsification by choosing random elements or $top_k$ elements, some information will be lost in aggregating gradients, and consequently, the quality of the model will be degraded. To overcome this problem, an elegant idea is proposed in [44] to use memory for tracking the history of entries and avoid the accumulation of compression errors. Similarly, we will employ a memory of aggregating gradients in order to compensate for the loss of information from sparsification. This is in addition to the local gradient tracking we incorporated in `FedCOMGATE`, however, despite the server control variate in SCAFFOLD, this memory is updated locally and is not required to be communicated to the server. We denote the memory in each client $j$ at round $r$ with $\boldsymbol{\nu}_j^{(r)}$. Thus, in Algorithm 2, we first need to compress the gradients added by the memory, using the $top_k$ operator as:

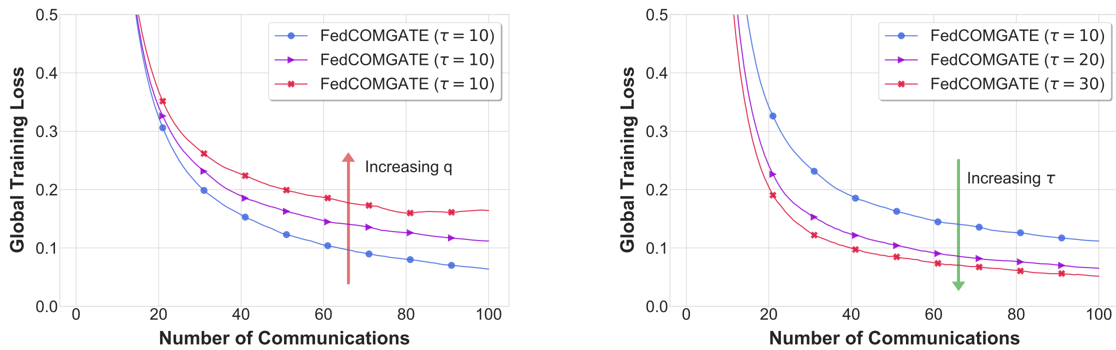
$$\boldsymbol{\Delta}_{j,s}^{(r)} = top_k \left\{ \left( \boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)} \right) + \boldsymbol{\nu}_j^{(r)} \right\} \tag{7}$$

Then, we will send this to the server for aggregation, where the server decompresses them, takes the average, and sends $\boldsymbol{\Delta}^{(r)}$ back to the clients. Each client updates its gradient tracking parameter as in Algorithm 2. Also, in this case, we need to update the memory parameter as:

$$\boldsymbol{\nu}_j^{(r+1)} = \boldsymbol{\nu}_j^{(r)} + \frac{1}{m} \left( \boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)} \right) - \boldsymbol{\Delta}^{(r)}, \tag{8}$$

where it keeps track of what was not captured by the aggregation using the sparsified gradients. Note that, unlike the quantized `FedCOMGATE`, in this approach, we cannot compress the downlink gradient broadcasting. However, since the cost of broadcasting is much lower than the uplink communication, this is negligible, especially in lower compression rates compared to quantized `FedCOMGATE`.

To show how the `FedCOMGATE` using sparsification with memory works in practice we will apply it to `MNIST` and `Fashion MNIST` datasets. Both of them are applied to an MLP model with two hidden layers, each with 200 neurons. For this experiment, we use the compression ratio parameter of $C_r$, which is the ratio between the size of communication in the compressed and without compression versions. Figure 6 shows the result of this algorithm by changing the compression rate. As it was observed by [45], in some compression




(a) Effect of quantization noise $q$

(b) Effect of local computation $\tau$

Figure 7: Investigating the effects of quantization noise and local computations on the convergence rate. We run the experiments on the `MNIST` dataset with a similar MLP model as before. In (a), we increase the noise of quantization by increasing the range of noise added to the zero-point of the quantizer operator. Increasing $q$ can degrade the convergence rate of the model. On the other hand, in (b), with the same level of quantization noise, we can increase the number of local computations $\tau$ to diminish the effects of quantization.

rates we can have similar or slightly better results than the without compression distributed SGD solution (here `FedGATE`), due to the use of memory. However, to gain more from the speedup and decreasing the compression rate, we will incur a residual error, as it can be seen in the results for the compression rate of 0.1.

**Effect of local computations.** Finally, we will show the effect of noise in quantization, characterized as $q$ in the paper, on the convergence rate, and how to address it. As it can be inferred from our theoretical analysis, increasing the noise of quantization would degrade the convergence rate of the model. This pattern can be seen in Figure 7(a) for the `MNIST` dataset, where we add noise to quantized arrays by adding a random integer to the zero-point of the quantization operator. By increasing the range of this noise, we can see that the convergence is getting worse with the same number of local computations. On the other hand, based on our analysis, we know that increasing the number of local computations will compensate for the quantization noise, which helps us to achieve the same results with lower communication rounds. This pattern is depicted in Figure 7(b), where we keep the quantization noise constant and increase the number of local computations.

# 7 Conclusion

In this paper we introduced a set of algorithms for federated learning which lower the communication overhead by periodic averaging and exchanging compressed signals. We considered two separate settings: (i) homogeneous setting in which all the probability distributions and loss functions are identical; and (ii) heterogeneous setting wherein the users' distributions and loss functions could be different. For both cases, we showed that our proposed methods both theoretically and numerically require less communication rounds between server and users compared to state-of-the-art federated algorithms that use compression.

# Acknowledgment


The authors would like to thank Amirhossein Reisizadeh for his comments on the first draft of the paper. We also gratefully acknowledge the generous support of NVIDIA for providing GPUs for our research.

# Appendix

The outline of our supplementary material follows. In Section A, we first elaborate further on related studies in the literature. In Section B, we propose variations of Algorithm 2 used in the experimental setup. Then, we present the proofs of our main theoretical results presented in the main body of the paper. In Section D, we present the convergence properties of our `FedCOM` method presented in Algorithm 1 for the **homogeneous** setting. In Section E, we present the convergence properties of our `FedCOMGATE` method presented in Algorithm 2 for the **heterogeneous** setting. In Section F, we present the proof of some of our intermediate lemmas.

## Table of Contents

# A Additional Related Work

In this section, we summarize and discuss additional related work. We separate the related work into two broad categories below.

**Local computation with periodic communication.** An elegant idea to reduce the number of communications in vanilla synchronous SGD is to perform averaging periodically instead of averaging models in all clients at every iteration [57], also known as local SGD. The seminal work of [44] was among the first to analyze the convergence of local SGD in the homogeneous setting and demonstrated that the number of communication rounds can be significantly reduced for smooth and strongly convex objectives while achieving linear speedup. This result is further improved in follow up studies [10, 11, 20, 46, 49, 56]. In [49], the error-runtime trade-off of local SGD is analyzed and it has been shown that it can also alleviate the synchronization delay caused by slow workers. From a practical viewpoint, few recent efforts explored adaptive communication strategies to communicate more frequently early in the process [10, 28, 40].

The analysis of local SGD in the heterogeneous setting, also known as federated averaging (FedAvg) [33], has seen a resurgence of interest very recently. While it is still an active research area [53], a few number of recent studies made efforts to understand the convergence of local SGD in a heterogeneous setting [12, 19, 20, 22, 24, 59]. Also, the personalization of local models for a better generalization in a heterogeneous setting is of great importance from both theoretical and practical point of view [7, 8, 27, 32, 43].

**Distributed optimization with compressed communication.** Another parallel direction of research has focused on reducing the size of communication by compressing the communicated messages. In quantization based methods, e.g. [4, 30, 41, 46], a quantization operator is applied before transmitting the gradient to server. A gradient acceleration approach with compression is proposed in [25]. In heterogeneous data distribution, [16] proposed the use of sign based SGD algorithms and [37] employed a quantization scheme in FedAvg with provable guarantees. In sparsification based methods, the idea is to transmit a smaller gradient vector by keeping only very few coordinates of local stochastic gradients, e.g., most significant entries [1, 31]. For these methods, theoretical guarantees have been provided in a few recent efforts [3, 45, 48, 54]. Note that, most of these studies rely on an error compensation technique as we employ in our experiments. We note that sketching methods are also employed to reduce the number of communication in [15].

The aforementioned studies mostly fall into the centralized distribution optimization. Recently a few attempts are made to explore the compression schema in a decentralized setting where each device shares compressed messages with direct neighbors over the underlying communication network [21, 38, 39, 42]. Another interesting direction for the purpose of reducing the communication complexity is to exploit the sparsity of communication network as explored in [12, 22, 50].

Finally, more thorough related works that study federated learning from different perspectives can be found in [17] and [23].

# B Variations of Algorithm 2

In this section we describe the details of variants of Algorithm 2 that are used in experiments.

**Without compression.** In this part, we first elaborate on a variant of Algorithm 2 without any compression involved, which we call it Federated Averaging with Local Gradient Tracking, `FedGATE`. Algorithm 3 describes the steps of `FedGATE`, which involves a local gradient tracking step. This algorithm is similar to the SCAFFOLD [19], however, the main difference is that we do not use any server control variate. In fact, `FedGATE`, as well as `FedCOMGATE`, are implicitly controlling the variance of the server model by controlling its subsidiaries' variances in local models. Therefore, there is no need to have another variable for this purpose, which can help us to greatly reduce the communication size, to half of what SCAFFOLD is using. Hence, even in the simple algorithm of `FedGATE`, we can gain the same convergence rate as SCAFFOLD, while enjoying the $2\times$ speedup in the communication. Note that, since the communication time of broadcasting from server

**Algorithm 3:** `FedGATE`$(R, \tau, \eta, \gamma)$ Federated Averaging with Local Gradient Tracking

**Inputs:** Number of communication rounds $R$, number of local updates $\tau$, learning rates $\gamma$ and $\eta$, initial global model $\boldsymbol{w}^{(0)}$, initial gradient tracking $\delta_j^{(0)} = \mathbf{0}$, $\forall j \in [m]$

**for** $r = 0, \ldots, R-1$ **do**
  **for** *each client* $j \in [m]$ **do in parallel**
    Set $\boldsymbol{w}_j^{(0,r)} = \boldsymbol{w}^{(r)}$
    **for** $c = 0, \ldots, \tau - 1$ **do**
      Set $\tilde{\boldsymbol{d}}_j^{(c,r)} = \tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}$ where $\tilde{\mathbf{g}}_j^{(c,r)} \triangleq \nabla f_j(\boldsymbol{w}_j^{(c,r)}; \mathcal{Z}_j^{(c,r)})$
      $\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\, \tilde{\boldsymbol{d}}_j^{(c,r)}$
    **end**
    Device **sends** $\boldsymbol{u}_j^{(r)} = \boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)}$ back to the server and gets $\boldsymbol{u}^{(r)}$
    Device **computes** $\bar{\boldsymbol{w}}^{(r)} = \boldsymbol{w}^{(r)} - \boldsymbol{u}^{(r)}$
    Device **updates** $\delta_j^{(r+1)} = \delta_j^{(r)} + \frac{1}{\eta\tau}\left(\bar{\boldsymbol{w}}^{(r)} - \boldsymbol{w}_j^{(\tau,r)}\right)$
    Device **updates** server model $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \gamma \boldsymbol{u}^{(r)}$ `// Option I`
  **end**
  Server **computes** $\boldsymbol{u}^{(r)} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_j^{(r)}$ and **broadcasts** back to clients
  Server **updates** $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \gamma \boldsymbol{u}^{(r)}$
  Server **broadcasts** $\boldsymbol{w}^{(r+1)}$ to all devices `// Option II`
**end**

to clients (or downlink communication) is negligible compared to gathering from clients to the server (or uplink communication), the overall communication complexity of this algorithm is close to FedAvg, and half of the SCAFFOLD, as it is depicted in Figure 3. Also, the communication complexity of `FedCOMGATE` is close to that of FedPAQ [37].

The common approach in federated learning without sampling for `FedGATE` and `FedCOMGATE` would be similar to `Option II` in Algorithm 3, where the server updates its model and broadcasts it to clients. This approach has one extra downlink step, that is negligible compared to the uplink steps, as it was mentioned before. However, when there is no sampling of the clients, we can avoid this extra downlink by using the `Option I`, where each local device keeps track of the server model and updates it based on what it gets for updating the gradient tracking variable. In practice, when sampling is not involved, we use `Option I`. In Section 6, we compare the performance of `FedGATE` and SCAFFOLD.

**User sampling.** One important aspect of federated learning is the sampling of clients since they might not be available all the time. Also, sampling clients can further reduce the per round communication complexity by aggregating information from a subset of clients instead of all clients. Hence, in Algorithm 4, we incorporate the sampling mechanism into our proposed `FedCOMGATE` algorithm. Based on this algorithm, at each communication round, the server selects a subset of clients $\mathcal{S}^{(r)} \subseteq [m]$, and sends the global server model only to selected devices in $\mathcal{S}^{(r)}$. The remaining steps of the algorithm are similar to Algorithm 2. In Section 6, we also study the effect of user sampling on the performance of `FedGATE`, `FedCOMGATE`, and other state-of-the-art methods for federated learning.

**Algorithm 4:** `FedCOMGATE`$(R, \tau, \eta, \gamma, k)$, `FedCOMGATE` algorithm with sampling of clients

**Inputs:** Number of communication rounds $R$, number of local updates $\tau$, learning rates $\gamma$ and $\eta$, initial global model $\boldsymbol{w}^{(0)}$, participation ratio of clients $k \in (0, 1]$, initial gradient tracking $\delta_j^{(0)} = \mathbf{0}$, $\forall j \in [m]$

**for** $r = 0, \ldots, R-1$ **do**
  Sever **selects** a subset of devices $\mathcal{S}^{(r)} \subseteq [m]$, with the size $\lfloor km \rfloor$
  Server **broadcasts** $w^{(r)}$ to the selected devices $j \in \mathcal{S}^{(r)}$
  **for** *each client* $j \in \mathcal{S}^{(r)}$ **do in parallel**
    Set $\boldsymbol{w}_j^{(0,r)} = \boldsymbol{w}^{(r)}$
    **for** $c = 0, \ldots, \tau - 1$ **do**
      Set $\tilde{\boldsymbol{d}}_{j,q}^{(c,r)} = \tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}$ where $\tilde{\mathbf{g}}_j^{(c,r)} \triangleq \nabla f_j(\boldsymbol{w}_j^{(c,r)}; \mathcal{Z}_j^{(c,r)})$
      $\boldsymbol{w}_j^{(c+1,r)} = \boldsymbol{w}_j^{(c,r)} - \eta\, \tilde{\boldsymbol{d}}_{j,q}^{(c,r)}$
    **end**
    Device **sends** $\boldsymbol{\Delta}_{j,q}^{(r)} = Q((\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)})/\eta)$ back to the server and gets $\boldsymbol{\Delta}_q^{(r)}$
    Device **updates** $\delta_j^{(r+1)} = \delta_j^{(r)} + \frac{1}{\tau}(\boldsymbol{\Delta}_{j,q}^{(r)} - \boldsymbol{\Delta}_q^{(r)})$
  **end**
  Server **computes** $\boldsymbol{\Delta}_q^{(r)} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\Delta}_{j,q}^{(r)}$ and **broadcasts** back to devices $j \in \mathcal{S}^{(r)}$
  Server **computes** $\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma\boldsymbol{\Delta}_q^{(r)}$
**end**

# C Some Definitions and Notation

Before stating our proofs we first formally define Polyak-Łojasiewicz and strongly convex functions.

**Assumption 6** (Polyak-Łojasiewicz)**.** *A function $f(\boldsymbol{w})$ satisfies the Polyak-Łojasiewicz condition with constant $\mu$ if $\frac{1}{2}\|\nabla f(\boldsymbol{w})\|_2^2 \geq \mu\big(f(\boldsymbol{w}) - f(\boldsymbol{w}^*)\big)$, $\forall \boldsymbol{w} \in \mathbb{R}^d$ with $\boldsymbol{w}^*$ is an optimal solution.*

**Assumption 7** ($\mu$-strong convexity)**.** *A function $f$ is $\mu$-strongly convex if it satisfies $f(\boldsymbol{u}) \geq f(\boldsymbol{v}) + \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v}\rangle + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2$, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$.*

We also introduce some notation for the clarity in presentation of proofs. Recall that we use $\mathbf{g}_i = \nabla f_i(\boldsymbol{w}) \triangleq \nabla f_i(\boldsymbol{w}; \mathcal{S}_i)$ and $\tilde{\mathbf{g}}_i \triangleq \nabla f(\boldsymbol{w}; \mathcal{Z}_i)$ for $1 \leq i \leq m$ to denote the full gradient and stochastic gradient at $i$th data shard, respectively, where $\mathcal{Z}_i \subseteq \mathcal{S}_i$ is a uniformly sampled mini-bath. The corresponding quantities evaluated at $i$th machine's local solution at $t$th iteration of optimization $\boldsymbol{w}_i^{(t)}$ are denoted by $\mathbf{g}_i^{(t)}$ and $\tilde{\mathbf{g}}_i^{(t)}$, where we abuse the notation and use $t = r\tau + c$ to denote the $c$th local update at $r$th round, i.e. $(c, r)$. We also define the following notations

$$\boldsymbol{w}^{(t)} = \{\boldsymbol{w}_1^{(t)}, \ldots, \boldsymbol{w}_m^{(t)}\},$$
$$\xi^{(t)} = \{\xi_1^{(t)}, \ldots, \xi_m^{(t)}\},$$

to denote the set of local solutions and sampled mini-batches at iteration $t$ at different machines, respectively. Finally, we use notation $\mathbb{E}[\cdot]$ to denote the conditional expectation $\mathbb{E}_{\xi^{(t)}|\boldsymbol{w}^{(t)}}[\cdot]$.

# D Results for the Homogeneous Setting

In this section, we study the convergence properties of our `FedCOM` method presented in Algorithm 1. Before stating the proofs for `FedCOM` in the homogeneous setting, we first mention the following intermediate lemmas.

**Lemma D.1.** *Under Assumptions 2 and 3, we have the following bound:*

$$\mathbb{E}_{Q,\xi^{(r)}}\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right] = \mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right] \leq \tau(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\left[\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2+\sigma^2\right] \tag{9}$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}_{\xi^{(r)}|\boldsymbol{w}^{(r)}}\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m}Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\|^2\right]\\
&=\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m}\underbrace{Q\left(\overbrace{\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}}^{\tilde{\mathbf{g}}_j^{(r)}}\right)}_{\tilde{\mathbf{g}}_{Qj}^{(r)}}\|^2\right]\right]\\
&\overset{①}{=}\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\left[\|\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_{Qj}^{(r)}-\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_Q\left[\tilde{\mathbf{g}}_{Qj}^{(r)}\right]\|^2\right]+\|\mathbb{E}_Q\left[\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_{Qj}^{(r)}\right]\|^2\right]\right]\\
&\overset{②}{=}\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\frac{1}{m^2}\sum_{j=1}^{m}\left[\left\|\tilde{\mathbf{g}}_{Qj}^{(r)}-\tilde{\mathbf{g}}_j^{(r)}\right\|^2\right]\right]+\left\|\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_j^{(r)}\right\|^2\right]\\
&\overset{③}{\leq}\mathbb{E}_{\xi^{(r)}}\left[\frac{1}{m}\sum_{j=1}^{m}\left[\frac{q}{m}\left\|\tilde{\mathbf{g}}_j^{(r)}\right\|^2+\left\|\tilde{\mathbf{g}}_j^{(r)}\right\|^2\right]\right]\\
&=(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\xi^{(r)}}\left[\left\|\tilde{\mathbf{g}}_j^{(r)}\right\|^2\right]\\
&=(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\left[\mathbb{E}_{\xi^{(r)}}\left[\left\|\tilde{\mathbf{g}}_j^{(r)}-\mathbb{E}_{\xi^{(r)}}\left[\tilde{\mathbf{g}}_j^{(r)}\right]\right\|^2\right]+\left\|\mathbb{E}_{\xi^{(r)}}\left[\tilde{\mathbf{g}}_j^{(r)}\right]\right\|^2\right]\\
&=(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\left[\mathbb{E}_{\xi^{(r)}}\left[\left\|\tilde{\mathbf{g}}_j^{(r)}-\mathbf{g}_j^{(r)}\right\|^2\right]+\left\|\mathbf{g}_j^{(r)}\right\|^2\right]
\end{aligned} \tag{10}$$

where ① holds due to $\mathbb{E}\left[\|\mathbf{x}\|^2\right]=\mathrm{Var}[\mathbf{x}]+\|\mathbb{E}[\mathbf{x}]\|^2$, ② is due to $\mathbb{E}_Q\left[\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_{Qj}^{(r)}\right]=\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_j^{(r)}$ and ③ follows from Assumption 2.

Next we show that from Assumptions 4, we have

$$\mathbb{E}_{\xi^{(r)}}\left[\left[\|\tilde{\mathbf{g}}_j^{(r)}-\mathbf{g}_j^{(r)}\|^2\right]\right]\leq\tau\sigma^2 \tag{11}$$

To do so, note that

$$\mathbb{E}_{\xi^{(r)}}\left[\left\|\tilde{\mathbf{g}}_j^{(r)}-\mathbf{g}_j^{(r)}\right\|^2\right]\overset{①}{=}\mathbb{E}_{\xi^{(r)}}\left[\left\|\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right]\right\|^2\right]$$

$$
\begin{aligned}
&= \mathrm{Var}\left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)}\right) \\
&\overset{②}{=} \sum_{c=0}^{\tau-1} \mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(c,r)}\right) \\
&= \sum_{c=0}^{\tau-1} \mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)}\right\|^2\right] \\
&\overset{③}{\leq} \tau\sigma^2
\end{aligned} \tag{12}
$$

where in ① we use the definition of $\tilde{\mathbf{g}}_j^{(r)}$ and $\mathbf{g}_j^{(r)}$, in ② we use the fact that mini-batches are chosen in i.i.d. manner at each local machine, and ③ immediately follows from Assumptions 3.

Replacing $\mathbb{E}_{\xi^{(r)}}\left[\|\tilde{\mathbf{g}}_j^{(r)} - \mathbf{g}_j^{(r)}\|^2\right]$ in (10) by its upper bound in (11) implies that

$$
\mathbb{E}_{\xi^{(r)}|\boldsymbol{w}^{(r)}}\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\|^2\right] = (\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\left[\tau\sigma^2 + \left\|\mathbf{g}_j^{(r)}\right\|^2\right] \tag{13}
$$

Further note that we have

$$
\left\|\mathbf{g}_j^{(r)}\right\|^2 = \|\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\|^2 \leq \tau\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 \tag{14}
$$

where the last inequality is due to $\left\|\sum_{j=1}^{n}\mathbf{a}_i\right\|^2 \leq n\sum_{j=1}^{n}\|\mathbf{a}_i\|^2$, which together with (13) leads to the following bound:

$$
\mathbb{E}_{\xi^{(r)}|\boldsymbol{w}^{(r)}}\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\|^2\right] \leq \tau(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\left[\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + \sigma^2\right], \tag{15}
$$

and the proof is complete. □

**Lemma D.2.** *Under Assumption 1, and according to the* `FedCOM` *algorithm the expected inner product between stochastic gradient and full batch gradient can be bounded with:*

$$
-\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right] \leq \frac{1}{2}\eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2\|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2\right] \tag{16}
$$

*Proof.* We have:

$$
\begin{aligned}
&-\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\mathbb{E}_Q\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}_Q^{(r)}\right\rangle\right] \\
&= -\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right\rangle\right] \\
&= -\left\langle\nabla f(\boldsymbol{w}^{(r)}), \eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\tilde{\mathbf{g}}_j^{(c,r)}\right]\right\rangle \\
&= -\eta\sum_{c=0}^{\tau-1}\frac{1}{m}\sum_{j=1}^{m}\left\langle\nabla f(\boldsymbol{w}^{(r)}), \mathbf{g}_j^{(c,r)}\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{①}{=} \frac{1}{2}\eta \sum_{c=0}^{\tau-1} \frac{1}{m} \sum_{j=1}^{m} \left[ -\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \|\nabla f(\boldsymbol{w}^{(r)}) - \nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 \right] \\
&\overset{②}{\leq} \frac{1}{2}\eta \sum_{c=0}^{\tau-1} \frac{1}{m} \sum_{j=1}^{m} \left[ -\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2 \|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2 \right]
\end{aligned} \tag{17}
$$

where ① is due to $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$, and ② follows from Assumption 1. □

The following lemma bounds the distance of local solutions from global solution at $r$th communication round.

**Lemma D.3.** *Under Assumptions 3 we have:*

$$
\mathbb{E}\left[ \|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2 \right] \leq \eta^2 \tau \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \eta^2 \tau \sigma^2 \tag{18}
$$

*Proof.* Note that

$$
\begin{aligned}
\mathbb{E}\left[ \left\| \boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)} \right\|_2^2 \right] &= \mathbb{E}\left[ \left\| \boldsymbol{w}^{(r)} - \left( \boldsymbol{w}^{(r)} - \eta \sum_{k=0}^{c} \tilde{\mathbf{g}}_j^{(k,r)} \right) \right\|_2^2 \right] \\
&= \mathbb{E}\left[ \left\| \eta \sum_{k=0}^{c} \tilde{\mathbf{g}}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{①}{=} \mathbb{E}\left[ \left\| \eta \sum_{k=0}^{c} \left( \tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \left[ \left\| \eta \sum_{k=0}^{c} \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{②}{=} \eta^2 \sum_{k=0}^{c} \mathbb{E}\left[ \left\| \left( \tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + (c+1)\, \eta^2 \sum_{k=0}^{c} \left[ \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\leq \eta^2 \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \left( \tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \tau \eta^2 \sum_{k=0}^{\tau-1} \left[ \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{③}{\leq} \eta^2 \sum_{k=0}^{\tau-1} \sigma^2 + \tau \eta^2 \sum_{k=0}^{\tau-1} \left[ \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&= \eta^2 \tau \sigma^2 + \eta^2 \sum_{k=0}^{\tau-1} \tau \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2
\end{aligned} \tag{19}
$$

where ① comes from $\mathbb{E}\left[\mathbf{x}^2\right] = \mathrm{Var}\left[\mathbf{x}\right] + \left[\mathbb{E}\left[\mathbf{x}\right]\right]^2$ and ② holds because $\mathrm{Var}\left(\sum_{j=1}^{n} \mathbf{x}_j\right) = \sum_{j=1}^{n} \mathrm{Var}\left(\mathbf{x}_j\right)$ for i.i.d. vectors $\mathbf{x}_i$ (and i.i.d. assumption comes from i.i.d. sampling), and finally ③ follows from Assumption 3. □

## D.1 Main result for the non-convex setting

Now we are ready to present our result for the homogeneous setting. We first state and prove the result for the general nonconvex objectives.

**Theorem D.4** (Non-convex)**.** *For* `FedCOM`$(\tau, \eta, \gamma)$*, for all* $0 \leq t \leq R\tau - 1$*, under Assumptions 1 to 3, if the learning rate satisfies*

$$
1 \geq \tau^2 L^2 \eta^2 + \left( \frac{q}{m} + 1 \right) \eta \gamma L \tau \tag{20}
$$

*and all local model parameters are initialized at the same point* $\boldsymbol{w}^{(0)}$*, then the average-squared gradient after* $\tau$ *iterations is bounded as follows:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \frac{2\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)}{\eta\gamma\tau R}+\frac{L\eta\gamma\left(\frac{q}{m}+1\right)}{m}\sigma^2+L^2\eta^2\tau\sigma^2 \tag{21}$$

*where* $\boldsymbol{w}^{(*)}$ *is the global optimal solution with function value* $f(\boldsymbol{w}^{(*)})$*.*

*Proof.* Before proceeding to the proof of Theorem D.4, we would like to highlight that

$$\boldsymbol{w}^{(r)}-\ \boldsymbol{w}_j^{(\tau,r)}=\eta\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}. \tag{22}$$

From the updating rule of Algorithm 1 we have

$$\boldsymbol{w}^{(r+1)}=\boldsymbol{w}^{(r)}-\gamma\eta\left(\frac{1}{m}\sum_{j=1}^{m}Q\Big(\sum_{c=0,r}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\Big)\right)=\boldsymbol{w}^{(r)}-\gamma\left[\frac{\eta}{m}\sum_{j=1}^{m}Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\right] \tag{23}$$

In what follows, we use the following notation to denote the stochastic gradient used to update the global model at $r$th communication round

$$\tilde{\mathbf{g}}_Q^{(r)}\triangleq\frac{\eta}{m}\sum_{j=1}^{m}Q\left(\frac{\boldsymbol{w}^{(r)}-\ \boldsymbol{w}_j^{(\tau,r)}}{\eta}\right)=\frac{\eta}{m}\sum_{j=1}^{m}Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right).$$

and notice that $\boldsymbol{w}^{(r)}=\boldsymbol{w}^{(r-1)}-\gamma\tilde{\mathbf{g}}^{(r)}$.

Then using the Assumption 2 we have:

$$\mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]=\frac{1}{m}\sum_{j=1}\left[-\eta\mathbb{E}_Q\left[Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\right]\right]=\frac{1}{m}\sum_{j=1}\left[-\eta\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\right]\triangleq\tilde{\mathbf{g}}^{(r)} \tag{24}$$

From the $L$-smoothness gradient assumption on global objective, by using $\tilde{\mathbf{g}}^{(r)}$ in inequality (22) we have:

$$f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\leq-\gamma\big\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}^{(r)}\big\rangle+\frac{\gamma^2L}{2}\|\tilde{\mathbf{g}}^{(r)}\|^2 \tag{25}$$

By taking expectation on both sides of above inequality over sampling, we get:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_Q\left[f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\right]\right]&\leq-\gamma\mathbb{E}\left[\mathbb{E}_Q\left[\big\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}_Q^{(r)}\big\rangle\right]\right]+\frac{\gamma^2L}{2}\mathbb{E}\left[\mathbb{E}_Q\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right]\\&\overset{(a)}{=}-\gamma\underbrace{\mathbb{E}\left[\left[\big\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}^{(r)}\big\rangle\right]\right]}_{\text{(I)}}+\frac{\gamma^2L}{2}\underbrace{\mathbb{E}\left[\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right]\right]}_{\text{(II)}}\end{aligned} \tag{26}$$

We proceed to use Lemma D.1, Lemma D.2, and Lemma D.3, to bound terms (I) and (II) in right hand side of (26), which gives

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}_Q\left[f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\right]\right]\\&\leq\gamma\frac{1}{2}\eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[-\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2-\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2+L^2\eta^2\sum_{c=0}^{\tau-1}\left[\tau\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2+\sigma^2\right]\right]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{(\frac{q}{m}+1)\gamma^2 L}{2}\left[\frac{\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + \frac{\tau\eta^2\sigma^2}{m}\right] \\
&\overset{①}{\leq} \frac{\gamma\eta}{2m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[-\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 - \left\|\mathbf{g}_j^{(c,r)}\right\|_2^2 + \tau L^2\eta^2\left[\tau\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2 + \sigma^2\right]\right] \\
&+ \frac{\gamma^2 L(\frac{q}{m}+1)}{2}\left[\frac{\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + \frac{\tau\eta^2\sigma^2}{m}\right] \\
&= -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \\
&- \left(1-\tau L^2\eta^2\tau - (\frac{q}{m}+1)\eta\gamma L\tau\right)\frac{\eta\gamma}{2m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(\frac{q}{m}+1)\right)\sigma^2 \\
&\overset{②}{\leq} -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(\frac{q}{m}+1)\right)\sigma^2
\end{aligned} \tag{27}
$$

where in ① we incorporate outer summation $\sum_{c=0}^{\tau-1}$, and ② follows from condition

$$
1 \geq \tau L^2\eta^2\tau + (\frac{q}{m}+1)\eta\gamma L\tau. \tag{28}
$$

Summing up for all $R$ communication rounds and rearranging the terms gives:

$$
\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right)}{\eta\gamma\tau R} + \frac{L\eta\gamma(\frac{q}{m}+1)}{m}\sigma^2 + L^2\eta^2\tau\sigma^2 \tag{29}
$$

From above inequality, is it easy to see that in order to achieve a linear speed up, we need to have $\eta\gamma = O\left(\frac{\sqrt{m}}{\sqrt{R\tau}}\right)$. □

**Corollary D.5** (Linear speed up)**.** *In Eq. (21) for the choice of* $\eta\gamma = O\left(\frac{1}{L}\sqrt{\frac{m}{R\tau(\frac{q}{m}+1)}}\right)$*, and* $\gamma \geq m$ *the convergence rate reduces to:*

$$
\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq O\left(\frac{L\sqrt{\left(\frac{q}{m}+1\right)}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{*})\right)}{\sqrt{mR\tau}} + \frac{\left(\sqrt{\left(\frac{q}{m}+1\right)}\right)\sigma^2}{\sqrt{mR\tau}} + \frac{m\sigma^2}{R\gamma^2}\right). \tag{30}
$$

*Note that according to Eq. (30), if we pick a fixed constant value for* $\gamma$*, in order to achieve an* $\epsilon$*-accurate solution,* $R = O\left(\frac{1}{\epsilon}\right)$ *communication rounds and* $\tau = O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)$ *local updates are necessary. We also highlight that Eq. (30) also allows us to choose* $R = O\left(\frac{\frac{q}{m}+1}{\epsilon}\right)$ *and* $\tau = O\left(\frac{1}{m\epsilon}\right)$ *to get the same convergence rate.*

**Remark 6.** *Condition in Eq. (20) can be rewritten as*

$$
\begin{aligned}
\eta &\leq \frac{-\gamma L\tau\left(\frac{q}{m}+1\right) + \sqrt{\gamma^2\left(L\tau\left(\frac{q}{m}+1\right)\right)^2 + 4L^2\tau^2}}{2L^2\tau^2} \\
&= \frac{-\gamma L\tau\left(\frac{q}{m}+1\right) + L\tau\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2 + 4}}{2L^2\tau^2} \\
&= \frac{\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2 + 4} - \left(\frac{q}{m}+1\right)\gamma}{2L\tau}
\end{aligned} \tag{31}
$$

*So based on Eq. (31), if we set* $\eta = O\left(\frac{1}{L\gamma}\sqrt{\frac{m}{R\tau\left(\frac{q}{m}+1\right)}}\right)$, *it implies that:*

$$R \geq \frac{\tau m}{\left(\frac{q}{m}+1\right)\gamma^2\left(\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2+4}-\left(\frac{q}{m}+1\right)\gamma\right)^2} \tag{32}$$

*We note that* $\gamma^2\left(\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2+4}-\left(\frac{q}{m}+1\right)\gamma\right)^2 = \Theta(1) \leq 5$ *therefore even for* $\gamma \geq m$ *we need to have*

$$R \geq \frac{\tau m}{5\left(\frac{q}{m}+1\right)} = O\left(\frac{\tau m}{\left(\frac{q}{m}+1\right)}\right) \tag{33}$$

***Therefore, for the choice of*** $\tau = O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)$***, due to condition in Eq. (33), we need to have*** $R = O\left(\frac{1}{\epsilon}\right)$***. Similarly, we can have*** $R = O\left(\frac{\frac{q}{m}+1}{\epsilon}\right)$ ***and*** $\tau = O\left(\frac{1}{m\epsilon}\right)$***.***

**Corollary D.6** (Special case, $\gamma = 1$)**.** *By letting* $\gamma = 1$, $q = 0$ *the convergence rate in Eq. (21) reduces to*

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \frac{2\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)}{\eta R\tau} + \frac{L\eta}{m}\sigma^2 + L^2\eta^2\tau\sigma^2 \tag{34}$$

*which matches the rate obtained in [49]. In this case the communication complexity and the number of local updates become*

$$R = O\left(\frac{m}{\epsilon}\right), \quad \tau = O\left(\frac{1}{\epsilon}\right). \tag{35}$$

*This simply implies that in this special case the convergence rate of our algorithm reduces to the rate obtained in [49], which indicates the tightness of our analysis.*

## D.2 Main result for the PL/strongly convex setting

We now turn to stating the convergence rate for the homogeneous setting under PL condition which naturally leads to the same rate for strongly convex functions.

**Theorem D.7** (PL or strongly convex)**.** *For* `FedCOM`$(\tau, \eta, \gamma)$*, for all* $0 \leq t \leq R\tau - 1$*, under Assumptions 1 to 3 and 6,if the learning rate satisfies*

$$1 \geq \tau^2 L^2\eta^2 + \left(\frac{q}{m}+1\right)\eta\gamma L\tau \tag{36}$$

*and if the all the models are initialized with* $\boldsymbol{w}^{(0)}$ *we obtain:*

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq (1-\eta\gamma\mu\tau)^R\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + \left(1+\frac{q}{m}\right)\frac{\gamma\eta L\sigma^2}{2m}\right] \tag{37}$$

*Proof.* From Eq. (27) under condition:

$$1 \geq \tau L^2\eta^2\tau + (\frac{q}{m}+1)\eta\gamma L\tau \tag{38}$$

we obtain:

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big] \leq -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(\frac{q}{m}+1)\right)\sigma^2$$

$$\leq -\eta\mu\gamma\tau\left(f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(r)})\right) + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(\frac{q}{m}+1)\right)\sigma^2 \tag{39}$$

which leads to the following bound:

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(*)})\Big] \leq (1-\eta\mu\gamma\tau)\Big[f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(*)})\Big] + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + (\frac{q}{m}+1)\gamma\right)\sigma^2 \tag{40}$$

By setting $\Delta = 1 - \eta\mu\gamma\tau$ we obtain the following bound:

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq \Delta^R\Big[f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big] + \frac{1-\Delta^R}{1-\Delta}\frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + (\frac{q}{m}+1)\gamma\right)\sigma^2 \\
&\leq \Delta^R\Big[f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big] + \frac{1}{1-\Delta}\frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + (\frac{q}{m}+1)\gamma\right)\sigma^2 \\
&= (1-\eta\mu\gamma\tau)^R\Big[f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big] + \frac{1}{\eta\mu\gamma\tau}\frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + (\frac{q}{m}+1)\gamma\right)\sigma^2
\end{aligned} \tag{41}$$

□

**Corollary D.8.** *If we let $\eta\gamma\mu\tau \leq \frac{1}{2}$, $\eta = \frac{1}{2L\left(\frac{q}{m}+1\right)\tau\gamma}$ and $\kappa = \frac{L}{\mu}$ the convergence error in Theorem D.7, with $\gamma \geq m$ results in:*

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq e^{-\eta\gamma\mu\tau R}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}\tau L^2\eta^2\sigma^2 + \left(1+\frac{q}{m}\right)\frac{\gamma\eta L\sigma^2}{2m}\right] \\
&\leq e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}L^2\frac{\tau\sigma^2}{L^2\left(\frac{q}{m}+1\right)^2\gamma^2\tau^2} + \left(1+\frac{q}{m}\right)\frac{L\sigma^2}{2\left(\frac{q}{m}+1\right)L\tau m}\right] \\
&= O\left(e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\sigma^2}{\left(\frac{q}{m}+1\right)^2\gamma^2\mu\tau} + \frac{\sigma^2}{\mu\tau m}\right) \\
&= O\left(e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\sigma^2}{\gamma^2\mu\tau} + \frac{\sigma^2}{\mu\tau m}\right)
\end{aligned} \tag{42}$$

*which indicates that to achieve an error of $\epsilon$, we need to have $R = O\left(\left(\frac{q}{m}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = \frac{1}{m\epsilon}$. Additionally, we note that if $\gamma \to \infty$, yet $R = O\left(\left(\frac{q}{m}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = \frac{1}{m\epsilon}$ will be necessary.*

## D.3 Main result for the general convex setting

**Theorem D.9** (Convex)**.** *For a general convex function $f(\boldsymbol{w})$ with optimal solution $\boldsymbol{w}^{(*)}$, using `FedCOM`$(\tau,\eta,\gamma)$ (Algorithm 1) to optimize $\tilde{f}(\boldsymbol{w},\phi) = f(\boldsymbol{w}) + \frac{\phi}{2}\|\boldsymbol{w}\|^2$, for all $0 \leq t \leq R\tau - 1$, under Assumptions 1 to 3, if the learning rate satisfies*

$$1 \geq \tau^2L^2\eta^2 + \left(\frac{q}{m}+1\right)\eta\gamma L\tau \tag{43}$$

*and if the all the models initiate with $\boldsymbol{w}^{(0)}$, with $\phi = \frac{1}{\sqrt{m\tau}}$ and $\eta = \frac{1}{2L\gamma\tau\left(1+\frac{q}{m}\right)}$ we obtain:*

$$\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] &\leq e^{-\frac{R}{2L\left(1+\frac{q}{m}\right)\sqrt{m\tau}}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \left[\frac{\sqrt{m}\sigma^2}{8\sqrt{\tau}\gamma^2\left(1+\frac{q}{m}\right)^2} + \frac{\sigma^2}{4\sqrt{m\tau}}\right] + \frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned} \tag{44}$$

We note that above theorem implies that to achieve a convergence error of $\epsilon$ we need to have $R = O\left(L\left(1+\frac{q}{m}\right)\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon}\right)$.

*Proof.* Since $\tilde{f}(\boldsymbol{w}^{(r)}, \phi) = f(\boldsymbol{w}^{(r)}) + \frac{\phi}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2$ is $\phi$-PL, according to Theorem D.7, we have:

$$\begin{aligned}
&\tilde{f}(\boldsymbol{w}^{(R)}, \phi) - \tilde{f}(\boldsymbol{w}^{(*)}, \phi)\\
&= f(\boldsymbol{w}^{(r)}) + \frac{\phi}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2 - \left(f(\boldsymbol{w}^{(*)}) + \frac{\phi}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2\right)\\
&\leq (1-\eta\gamma\phi\tau)^R\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\phi}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + \left(1+\frac{q}{m}\right)\frac{\gamma\eta L\sigma^2}{2m}\right]
\end{aligned} \tag{45}$$

Next rearranging Eq. (45) and replacing $\mu$ with $\phi$ leads to the following error bound:

$$\begin{aligned}
&f(\boldsymbol{w}^{(R)}) - f^*\\
&\leq (1-\eta\gamma\phi\tau)^R\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\phi}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + \left(1+\frac{q}{m}\right)\frac{\gamma\eta L\sigma^2}{2m}\right]\\
&\qquad + \frac{\phi}{2}\left(\|\boldsymbol{w}^*\|^2 - \left\|\boldsymbol{w}^{(r)}\right\|^2\right)\\
&\leq e^{-(\eta\gamma\phi\tau)R}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\phi}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + \left(1+\frac{q}{m}\right)\frac{\gamma\eta L\sigma^2}{2m}\right] + \frac{\phi}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned} \tag{46}$$

Next, if we set $\phi = \frac{1}{\sqrt{m\tau}}$ and $\eta = \frac{1}{2\left(1+\frac{q}{m}\right)L\gamma\tau}$, we obtain that

$$\begin{aligned}
&f(\boldsymbol{w}^{(R)}) - f^*\\
&\leq e^{-\frac{R}{2\left(1+\frac{q}{m}\right)L\sqrt{m\tau}}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \sqrt{m\tau}\left[\frac{\sigma^2}{8\tau\gamma^2\left(1+\frac{q}{m}\right)^2} + \frac{\sigma^2}{4\tau m}\right] + \frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2,
\end{aligned} \tag{47}$$

thus the proof is complete. ∎

# E Results for the Heterogeneous Setting

In this section, we study the convergence properties of `FedCOMGATE` method presented in Algorithm 2. For this algorithm recall that the update rule can be written as:

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma\frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right) = \boldsymbol{w}^{(r)} - \gamma\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\right) \tag{48}$$

Before stating the proofs for `FedCOMGATE` in the heterogeneous setting, we first mention the following intermediate lemmas.

**Lemma E.1.** *Under Assumptions 2, 4 and 5, for the updates of `FedCOMGATE` we have the following bound:*

$$\begin{aligned}\mathbb{E}&\left[\mathbb{E}_Q\left[\left\|\frac{\eta}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\right\|^2\right]\right]\\ &\leq (q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2 G_q\end{aligned} \tag{49}$$

*Proof.* First, note that the expression on the left hand side of (49) can be upper bounded by

$$\begin{aligned}
&\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\|^2\Big]\\
&\overset{①}{=} \eta^2\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\underbrace{Q\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\right)}_{\tilde{\mathbf{g}}_Q^{(r)}}\|^2 + G_q\Big]\\
&= \eta^2\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\underbrace{Q\left(\overbrace{\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)}\right)}^{\tilde{\mathbf{g}}^{(r)}}\right)}_{\tilde{\mathbf{g}}_Q^{(r)}}\|^2 + G_q\Big]\\
&=\eta^2\mathbb{E}_\xi\left[\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)} - \mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]\|^2\right] + \|\mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]\|^2\right] + \eta^2 G_q\\
&=\eta^2\mathbb{E}_\xi\left[\mathbb{E}_Q\left[\left\|\tilde{\mathbf{g}}_Q^{(r)} - \tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q\\
&\leq\eta^2\mathbb{E}_\xi\left[q\left\|\tilde{\mathbf{g}}^{(r)}\right\|^2 + \left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q\\
&= (q+1)\eta^2\mathbb{E}_\xi\left[\left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q\\
&= (q+1)\eta^2\mathbb{E}_\xi\left[\left\|\tilde{\mathbf{g}}^{(r)} - \mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2\right] + (q+1)\eta^2\left\|\mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2 + \eta^2 G_q
\end{aligned} \tag{50}$$

where ① comes from Assumption 2.

Moreover, under Assumption 4, we can show following variance bound from the averaged stochastic gradient:

$$\mathbb{E}\left[\left[\|\tilde{\mathbf{g}}^{(r)} - \mathbf{g}^{(r)}\|^2\right]\right] \leq \frac{\tau\eta^2\sigma^2}{m} \tag{51}$$

To prove this claim, note that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\right\|^2\right] &\overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\left[\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|^2\right] \\
&\overset{②}{=} \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\left[\left\|\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right]\right\|^2\right] \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right\|^2\right] \\
&\overset{③}{\leq} \frac{1}{m^2}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\sigma^2 \qquad (52) \\
&= \frac{\tau\sigma^2}{m} \qquad (53)
\end{aligned}
$$

where in ① we use the definition of $\tilde{\mathbf{g}}^t$ and $\mathbf{g}^t$, in ② we use the fact that mini-batches are chosen in i.i.d. manner at each local machine, and ③ immediately follows from Assumptions 4.

Now replace the upper bound in (51) into the last expression in (50) to obtain

$$
\begin{aligned}
&\mathbb{E}_\xi\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{\Delta}_j^{(r)}\right)\|^2\right] \\
&\leq (q+1)\eta^2\tau\frac{\sigma^2}{m}+(q+1)\eta^2\left\|\mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2+\eta^2 G_q \qquad (54)
\end{aligned}
$$

Next, note that i.i.d. data distribution implies $\mathbb{E}[\tilde{\mathbf{g}}_j^{(r)}]=\mathbf{g}_j^{(r)}$, from which we have

$$
\begin{aligned}
\left\|\mathbb{E}\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2 &= \|\mathbf{g}^{(r)}\|^2 \\
&\leq \left\|\frac{1}{m}\sum_{j=1}\left[\sum_{c=0}^{\tau-1}g_j^{(c,r)}\right]\right\|^2 \\
&\overset{①}{\leq} \left\|\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right\|^2 \\
&\overset{②}{\leq} \tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}g_j^{(c,r)}\right\|^2 \\
&= \tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 \qquad (55)
\end{aligned}
$$

where ① follows from convexity of $\|.\|$ and ② is due to $\left\|\sum_{j=1}^{n}\mathbf{a}_i\right\|^2\leq n\sum_{j=1}^{n}\|\mathbf{a}_i\|^2$.

Applying this upper bound into (54) implies that

$$
\mathbb{E}_\xi\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{\Delta}_j^{(r)}\right)\|^2\right]
$$

$$\leq (q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}^{(r)}\right\|^2 + \eta^2 G_q, \tag{56}$$

and the proof is complete. $\square$

**Lemma E.2.** *Under Assumptions 1, for the updates of* `FedCOMGATE` *we can show that the expected inner product between stochastic gradient and full batch gradient can be bounded as*

$$\begin{aligned}
&-\eta\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{w}^{(t)}), \tilde{\mathbf{g}}^{(t)}\right\rangle\right] \\
&\leq \frac{1}{2}\eta\sum_{c=0}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2\sum_{j=1}^{m}\frac{1}{m}\|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2\right]
\end{aligned} \tag{57}$$

*Proof.* This proof is relatively as we state in the following expressions:

$$\begin{aligned}
&-\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\mathbb{E}_Q\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right] \\
&= -\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right\rangle\right] \\
&= -\left\langle\nabla f(\boldsymbol{w}^{(r)}), \eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\tilde{\mathbf{g}}_j^{(c,r)}\right]\right\rangle \\
&= -\eta\sum_{c=0}^{\tau-1}\left\langle\nabla f(\boldsymbol{w}^{(r)}), \frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\rangle \\
&\overset{①}{=} \frac{1}{2}\eta\sum_{c=0}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \|\nabla f(\boldsymbol{w}^{(r)}) - \frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2\right] \\
&\overset{②}{\leq} \frac{1}{2}\eta\sum_{c=0}^{\tau-1}\frac{1}{m}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\frac{1}{m}\sum_{j=1}^{m}\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2\frac{1}{m}\sum_{j=1}^{m}\|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2\right],
\end{aligned} \tag{58}$$

where ① is due to $2\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2$, and ② follows from Assumption 1.

$\square$

**Lemma E.3.** *Under Assumptions 2, 4 and 5, with $30\eta^2L^2\tau^2 \leq 1$ we have:*

$$\begin{aligned}
&\frac{1}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 \\
&\quad\leq 36\eta^2\tau^2\sigma^2 + \frac{8\eta^2}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad + 10\eta^2(\eta\gamma)^2(q+1)L^2\left[\frac{\tau^4}{R}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r-1)}\right\|^2\right] + \tau^4\frac{\sigma^2}{m} + \tau^3G_q\right] \\
&\qquad + \frac{20\eta^2\tau^2}{R}\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2.
\end{aligned} \tag{59}$$

The proof of this intermediate lemma is deferred to Appendix F.

## E.1 Main result for the nonconvex setting

**Theorem E.4** (General Non-convex). *For* `FedCOMGATE`$(\tau, \eta, \gamma)$*, for all* $0 \le t \le R\tau - 1$*, under Assumptions 1, 2, 4 and 5 and if the learning rate satisfies*

$$1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - L\eta\gamma\tau(q+1) \ge 0 \quad \& \quad 30\eta^2L^2\tau^2 \le 1 \tag{60}$$

*and all local model parameters are initialized at the same point* $\bar{\boldsymbol{w}}^{(0)} = \boldsymbol{w}^{(0)}$*, we obtain:*

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 \le{}& \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right)}{\tau\eta\gamma R} + \frac{(q+1)\gamma L\eta\sigma^2}{m} + 36\eta^2L^2\tau\sigma^2 + 10\eta^2L^4\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}\\ &+ 10\eta^2L^4\tau^2(\eta\gamma)^2(q+1)G_q + \frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)}) - f_j(\boldsymbol{w}_j^{(*)})\right] + \frac{16\eta^3L^2\tau^2}{R}\sigma^2\\ &+ \frac{32\eta^2L^3\tau^2}{R}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\gamma\eta L}{\tau}G_q\end{aligned} \tag{61}$$

*Proof.* Before proceeding to the proof we need to review some properties of our algorithm:

1) $\delta_j^{(0)} = 0$

2) $\boldsymbol{\Delta}_{j,q}^{(r)} = \mathsf{Q}\left(\left(\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)}\right)/\eta\right)$

3) $\boldsymbol{\Delta}_q^{(r)} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{\Delta}_{j,q}^{(r)}$

4) $\delta_j^{(r)} = \frac{1}{\tau}\sum_{k=0}^{r}\left(\boldsymbol{\Delta}_q^{(k)} - \boldsymbol{\Delta}_{j,q}^{(k)}\right)$

5) $\frac{1}{m}\sum_{j=1}^m \delta_j^{(r)} = 0.$

6) We have:

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \gamma\eta\frac{1}{m}\sum_{j=1}^{m}Q\left(-\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jQ}^{(c,r)}\right) = \boldsymbol{w}^{(r)} - \gamma\eta\frac{1}{m}\sum_{j=1}^{m}Q\left(\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}\right]\right)$$

which is equivalent to the update rule of the global model of Algorithm 2.

7) We have:

$$\begin{aligned}\delta_j^{(r)} &= \delta_j^{(r-1)} + \frac{1}{\tau}\left(\boldsymbol{\Delta}_q^{(k)} - \boldsymbol{\Delta}_{j,q}^{(k)}\right)\\ &= \delta_j^{(r-1)} + \frac{1}{\tau}\left(\frac{1}{m}\sum_{j=1}^{m}Q\left(-\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \delta_j^{(r-1)}\right)\right) + Q\left(\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \delta_j^{(r-1)}\right)\right)\right)\end{aligned} \tag{62}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_Q\left[\delta_j^{(r)}\right] &= \frac{1}{\tau}\left(-\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)} + \sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right) + \frac{1}{\tau}\left(\frac{1}{m}\sum_{j=1}^{m}\delta_j^{(r-1)}\right)\\ &= \frac{1}{\tau}\left(-\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)} + \sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\end{aligned} \tag{63}$$

8) From item (7), for $R \geq 1$ we obtain:

$$\mathbb{E}_Q\left[\tilde{\boldsymbol{d}}_{jq}^{(c,r)}\right] = \mathbb{E}_Q\left[\tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}\right] = \tilde{\mathbf{g}}_j^{(c,r)} + \frac{1}{\tau}\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)} - \sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right) = \tilde{\boldsymbol{d}}_j^{(c,r)} \tag{64}$$

We would like to also highlight that

$$-\eta Q\left(\frac{\boldsymbol{w}^{(r)} - \ \boldsymbol{w}_j^{(\tau,r)}}{\eta}\right) = -\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jQ}^{(c,r)}\right) \tag{65}$$

Towards this end, recalling the notation

$$\tilde{\mathbf{g}}_Q^{(r)} \triangleq \frac{1}{m}\sum_{j=1}\left[-\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jQ}^{(c,r)}\right)\right],$$

and using the Assumption 2 we have:

$$\mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right] = \frac{1}{m}\sum_{j=1}\left[-\eta\mathbb{E}_Q\left[Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jq}^{(c,r)}\right)\right]\right] = \frac{1}{m}\sum_{j=1}\left[-\eta\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\right] \triangleq \tilde{\mathbf{g}}^{(r)} \tag{66}$$

Then following the $L$-smoothness gradient assumption on global objective, by using $\tilde{\mathbf{g}}^{(r)}$ in inequality (65) we have:

$$f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)}) \leq -\gamma\big\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}_Q^{(r)}\big\rangle + \frac{\gamma^2 L}{2}\|\tilde{\mathbf{g}}_Q^{(r)}\|^2 \tag{67}$$

By taking expectation on both sides of above inequality over sampling, we get:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_Q\left[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\right]\right] &\leq -\gamma\mathbb{E}\left[\mathbb{E}_Q\left[\big\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}_Q^{(r)}\big\rangle\right]\right] + \frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}_Q\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right] \\ &\overset{①}{=} -\gamma\underbrace{\mathbb{E}\left[\left[\big\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)}\big\rangle\right]\right]}_{\text{(I)}} + \frac{\gamma^2 L}{2}\underbrace{\mathbb{E}\left[\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right]\right]}_{\text{(II)}}\end{aligned} \tag{68}$$

where ① follows from Eq. (66). Next, by plugging back the results in Lemma E.1, Lemma E.2, and Lemma E.3 we obtain

$$\begin{aligned}&\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big] \\ &\leq \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\Big[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \frac{L^2}{m}\sum_{j=1}^{m}\left[\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2\right]\Big] \\ &+ \frac{\gamma^2 L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2 G_q\right]\end{aligned} \tag{69}$$

which leads to

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big]$$

$$
\begin{aligned}
&\leq -\frac{\gamma\eta}{2}\frac{\tau}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \frac{\gamma\eta}{2}\frac{1}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&\quad + \frac{\gamma\eta}{2}\frac{L^2}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 \\
&\quad + \frac{\gamma^2 L\eta^2\tau\sigma^2(q+1)}{2m} + \frac{\gamma^2\eta^2 L}{2}G_q + \frac{(q+1)\gamma^2 L\eta^2\tau}{2R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\|\frac{1}{m}\sum_{j=1}^{m}g_j^{(c,r)}\|^2 \\
&\overset{①}{\leq} -\frac{\gamma\eta}{2}\frac{\tau}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \frac{\gamma\eta}{2}\frac{1}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&\quad + \frac{\gamma\eta}{2}\frac{L^2}{R}36\eta^2\tau^2\sigma^2 + \frac{\gamma\eta}{2}\frac{8L^2\eta^2}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + \frac{L^2\gamma\eta}{2}10\eta^2(\eta\gamma)^2(q+1)L^2\left[\frac{\tau^4}{R}\sum_{r=0}^{R-1}\sum_{c=0,r=0}^{\tau-1}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2\right] + \tau^4\frac{\sigma^2}{m} + \tau^3 G_q\right] \\
&\quad + \frac{L^2\gamma\eta}{2}\frac{20\eta^2\tau^2}{R}\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \\
&\quad + \frac{\gamma^2\eta^2 L}{2}G_q + \frac{(q+1)\gamma^2 L\eta^2\tau\sigma^2}{2m} + \frac{(q+1)\gamma^2 L\eta^2\tau}{2R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\|\frac{1}{m}\sum_{j=1}^{m}g_j^{(c,r)}\|^2 \\
&= -\left(\frac{\gamma\eta}{2}\frac{\tau}{R} - \frac{L^2\gamma\eta}{2}\frac{20\eta^2\tau^3}{R}\right)\sum_{r=0}^{R-1}\|\mathbf{g}^{(r)}\|_2^2 \\
&\quad - \frac{\gamma\eta}{2}\left(1 - L^2 10\eta^2(\eta\gamma)^2(q+1)L^2\tau^4 - L(q+1)\eta\gamma\tau\right)\frac{1}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&\quad + \frac{\gamma\eta}{2}\frac{L^2}{R}36\eta^2\tau^2\sigma^2 + \frac{\gamma\eta}{2}\frac{8L^2\eta^2}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + \frac{L^2\gamma\eta}{2}10\eta^2(\eta\gamma)^2(q+1)L^2\left[\tau^4\frac{\sigma^2}{m} + \tau^3 G_q\right] + \frac{(q+1)\gamma^2 L\eta^2\tau\sigma^2}{2m} + \frac{\gamma^2\eta^2 L}{2}G_q \\
&\overset{②}{\leq} -\quad \frac{\gamma\eta}{2}\frac{\tau}{R}\left(1 - L^2 20\eta^2\tau^2\right)\sum_{r=0}^{R-1}\|\mathbf{g}^{(r)}\|_2^2 \\
&\quad + \frac{\gamma\eta}{2}\frac{L^2}{R}36\eta^2\tau^2\sigma^2 + \frac{\gamma\eta}{2}\frac{8L^2\eta^2}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + \frac{L^2\gamma\eta}{2}10\eta^2(\eta\gamma)^2(q+1)L^2\left[\tau^4\frac{\sigma^2}{m} + \tau^3 G_q\right] + \frac{(q+1)\gamma^2 L\eta^2\tau\sigma^2}{2m} + \frac{\gamma^2\eta^2 L}{2}G_q
\end{aligned} \tag{70}
$$

where ① comes from Lemma E.3 and ② follows by imposing the following condition:

$$
1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - (q+1)L\eta\gamma\tau \geq 0. \tag{71}
$$

Rearranging Eq. (70) we obtain:

$$
\left(1 - 20\eta^2 L^2\tau^2\right)\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2
$$

$$\leq \frac{2\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{*})\right)}{\tau\eta\gamma R}+\frac{(q+1)\gamma L\eta\sigma^2}{m}+36\eta^2L^2\tau\sigma^2+10\eta^2L^4\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}$$
$$+10\eta^2L^4\tau^2(\eta\gamma)^2(q+1)G_q+\underbrace{\frac{8\eta^2L^2}{m\tau R}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2}_{\text{(IV)}}+\frac{\gamma\eta L}{\tau}G_q, \tag{72}$$

and the claim follows.

The final step is to simplify the term (IV). To this purpose, first notice that

$$\begin{aligned}
&\frac{8\eta^2L^2}{m\tau R}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&=\frac{8\eta^2L^2\tau}{m\tau R}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&\overset{①}{\leq}\frac{8\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2\\
&\leq\frac{16\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[\left\|\mathbf{g}_j^{(c,0)}\right\|^2+\left\|\mathbf{g}^{(0)}\right\|^2\right]\\
&=\frac{16\eta^2L^2\tau^2}{mR}\sum_{j=1}^{m}\left[\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)}\right\|^2\right]+\frac{16\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{②}{\leq}\frac{16\eta^2L^2\tau^2}{mR}\sum_{j=1}^{m}\left[\frac{2\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]}{\eta\tau}+\eta\sigma^2\right]+\frac{16\eta^2L^2\tau^2}{R}\left\|\mathbf{g}^{(0)}\right\|^2\\
&=\frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]+\frac{16\eta^3L^2\tau^2}{R}\sigma^2+\frac{16\eta^2L^2\tau^2}{R}\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{③}{\leq}\frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]+\frac{16\eta^3L^2\tau^2}{R}\sigma^2+\frac{32\eta^2L^3\tau^2}{R}\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)
\end{aligned}\tag{73}$$

where ① comes from $\|\sum_{i=1}^{n}\mathbf{a}_i\|^2\leq n\sum_{i=1}^{n}\|\mathbf{a}_i\|^2$, in ② we used the standard convergence proof of gradient descent for non-convex objectives [6], where $\boldsymbol{w}_j^{(*)}$ is the local minimizer of objective function $f_j(.)$, and finally, ③ follows from (smoothness assumption) inequality $\left\|\mathbf{g}^{(0)}\right\|^2\leq 2L\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)$ (see [9, 46] for more details). This completes the proof.

□

**Remark 7.** *If we let $\eta\gamma=\frac{1}{L}\sqrt{\frac{m}{R\tau(q+1)}}$, and want to make sure that the condition in Eq. (60) is satisfied simultaneously, we need to have*

$$1\geq\frac{10m^2\tau^2}{\gamma^2R^2(q+1)}+\sqrt{\frac{m\tau\,(q+1)}{R}} \tag{74}$$

*This inequality is a polynomial of degree 4 with respect to $R$, therefore characterizing exact solution could be difficult. So, by letting $\gamma\geq\sqrt{20}m$ we derive an necessary solution here as follows:*

$$R\geq m\tau\left(\frac{q+1}{2}\right) \tag{75}$$

*We note that if we solve this inequality such as Eq. (32) we are expecting to degrade the dependency on $q$. This condition requires having $R = \left(\frac{q+1}{m\epsilon}\right)$ and $\tau = \left(\frac{1}{m\epsilon}\right)$.*

**Corollary E.5** (Linear speed up with fix global learning rate)**.** *Considering the condition $30\eta^2L^2\tau^2 \leq 1$, we have $1 - 20\eta^2L^2 = \Theta(1)$. Therefore, in Eq. (61) if we set $\eta\gamma = O\left(\frac{1}{L}\sqrt{\frac{m}{R\tau(q+1)}}\right)$, $\gamma \geq m$ leads to:*

$$
\begin{aligned}
&\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \\
&\leq O\Big(\frac{L\sqrt{q+1}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right)}{\sqrt{mR\tau}} + \frac{\sqrt{q+1}\sigma^2}{\sqrt{mR\tau}} + \frac{m\sigma^2}{R\,(q+1)\,\gamma^2} + \frac{m\sigma^2\tau}{(q+1)\,\gamma^2R^2} + \frac{m^2G_q}{(q+1)\,\gamma^2R^2} \\
&\quad + \frac{L\sqrt{\tau}}{\gamma\sqrt{q+1}\sqrt{m}R^{1.5}}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)}) - f_j(\boldsymbol{w}_j^{(*)})\right] + \frac{16m\sqrt{m}\sqrt{\tau}\sigma^2}{L\gamma^3R^2(q+1)\sqrt{R(q+1)}} + \frac{Lm}{\gamma^2R^2(q+1)}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{mG_q}{R\tau\gamma^2\,(q+1)}\Big),
\end{aligned}
$$

*then by letting $\gamma \geq m$ we improve the convergence rate of [19] and [36] with tuned global and local learning rates, showing that we can archive the error $\epsilon$ with $R = \Theta\left((1+q)\,\epsilon^{-1}\right)$ and $\tau = \Theta\left(\frac{1}{m\epsilon}\right)$, which matches the communication and computational complexity of [19] and [26], which shows that obtained rate is tight. We highlight that the communication complexity of our algorithm is better than [19] in terms of number of bits per iteration as we do not use additional control variable.*

**Remark 8.** *We note that the conditions in Eq. (60) can be rewritten as*

$$
1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - L(q+1)\eta\gamma\tau \geq 0 \quad \& \quad 30\eta^2L^2\tau^2 \leq 1 \tag{76}
$$

*which implies that the choice of $\eta \leq \frac{1}{L\gamma(q+1)\tau\sqrt{30}}$ satisfies both conditions for $\gamma \geq m$.*

## E.2 Main result for the PL/strongly convex setting

**Theorem E.6** (Strongly convex or PL)**.** *For* `FedCOMGATE`$(\tau, \eta, \gamma)$*, for all $0 \leq t \leq R\tau - 1$, under Assumptions 1, 2, 4, 5 and 6 and if the learning rate satisfies*

$$
1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - L\eta\gamma\tau(q+1) \geq 0 \quad \& \quad 30\eta^2L^2\tau^2 \leq 1 \tag{77}
$$

*and all local model parameters are initialized at the same point $\boldsymbol{w}^{(0)}$, we obtain:*

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq \left(1 - \frac{\mu\eta\gamma\tau}{3}\right)^R\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{3}{\mu}\Big[L^2 18\eta^2\tau\sigma^2 + \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\left\|\left(\boldsymbol{w}_j^{(0,0)} - \boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + 16L^3\tau^2\eta^2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big].
\end{aligned} \tag{78}
$$

*Proof.* To prove our claim we use the following lemma. The proof of this intermediate lemma is deferred to Appendix F.

**Lemma E.7.** *With $30\eta^2L^2\tau^2 \leq 1$, under Assumptions 1, 2, 4 and 5 we have:*

$$
\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2
$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\eta\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2 \\
&\le 36\eta^2\tau^2\sigma^2 + \frac{8\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + 10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r-1)}\right\|^2 \\
&\quad + 10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + 20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2
\end{aligned} \tag{79}
$$

Now we proceed to prove the claim of Theorem E.6. Note that

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big] \\
&\le \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \frac{L^2}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2\right] \\
&\quad + \frac{\gamma^2L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2G_q\right] \\
&= -\frac{\tau\gamma\eta}{2}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 + \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\left[-\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \frac{L^2}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2\right] \\
&\quad + \frac{\gamma^2L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2G_q\right]
\end{aligned} \tag{80}
$$

which leads to the following:

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big] \\
&= -\frac{\tau\gamma\eta}{2}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&\quad + \frac{1}{2}\gamma\eta L^2\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 \\
&\quad + \frac{\gamma^2L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2G_q\right] \\
&\le -\frac{\tau\gamma\eta}{2}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&\quad + 18L^2\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + \gamma\eta5\eta^2L^4\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\Big[\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r-1)}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 5L^2\gamma\eta\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q\Big] + \gamma\eta 10\eta^2\tau^3L^2\left\|\mathbf{g}^{(r)}\right\|^2 \\
&+ \frac{\gamma^2 L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2 G_q\right] \\
=& -\left(\frac{\tau\gamma\eta}{2} - \gamma L^2\eta 10\eta^2\tau^3\right)\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 \\
&- \frac{1}{2}\gamma\eta\left(1 - \gamma\eta 10\eta^2L^4\tau^4(\eta\gamma)^2(q+1) - (q+1)L\eta\gamma\tau\right)\sum_{c=0,r}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 \\
&+ L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2 L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2} \\
\overset{①}{\leq}& -\left(\frac{\tau\gamma\eta}{2} - \gamma L^2\eta 10\eta^2\tau^3\right)\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 \\
&+ L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{\eta^2\gamma^2 L}{2}\frac{(q+1)\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2}
\end{aligned}
\tag{81}
$$

where ① follows from

$$
1 - \gamma\eta 10\eta^2L^4\tau^4(\eta\gamma)^2(q+1) - (q+1)L\eta\gamma\tau \geq 0 \tag{82}
$$

Next Eq. (81) leads us to

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq \left(1 - \mu\tau\gamma\eta + 20\mu\gamma\eta^3L^2\tau^3\right)\left(f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2 L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2} \\
&\overset{①}{=} \Delta^r\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{1-\Delta^r}{1-\Delta}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2 L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2}\Big] \\
&\leq \Delta^r\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{1}{1-\Delta}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2 L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2}\Big]
\end{aligned}
\tag{83}
$$

where ① holds because of $\Delta = 1 - \mu\tau\gamma\eta + 20\mu\gamma\eta^3 L^2\tau^3$ which leads the following bound:

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\quad \leq \left(1 - \mu\eta\gamma\tau\left(1 - 20\eta^2L^2\tau^2\right)\right)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\qquad + \frac{1}{\mu\eta\gamma\tau\left(1 - 20\eta^2L^2\tau^2\right)}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad\quad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2}\Big] \\
&\quad = \left(1 - \mu\eta\gamma\tau\left(1 - 20\eta^2L^2\tau^2\right)\right)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\qquad + \frac{1}{\mu\left(1 - 20\eta^2L^2\tau^2\right)}\Big[L^2 18\eta^2\tau\sigma^2 + \frac{4L^2\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big] \\
&\quad \overset{①}{\leq} \left(1 - \frac{\mu\eta\gamma\tau}{3}\right)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\qquad + \frac{3}{\mu}\Big[L^2 18\eta^2\tau\sigma^2 + \underbrace{L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2}_{\text{(V)}} \\
&\qquad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned}
\tag{84}
$$

where in ① we used the condition $30\eta^2L^2\tau^2 \leq 1$.

Finally we continue with bounding term (V):

$$
\begin{aligned}
&\frac{4L^2\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad = \frac{4L^2\eta^2\tau}{m\tau}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(*)} + \mathbf{g}_j^{(*)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad \leq \frac{8L^2\eta^2}{m}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(*)}\right)\right\|^2 + \frac{8L^2\eta^2}{m}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(*)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad = \frac{8L^2\eta^2}{m}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(*)}\right)\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)} - \mathbf{g}^{(0)}\right\|^2 \\
&\qquad \leq \frac{8L^2\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)} - \mathbf{g}^{(0)}\right\|^2 \\
&\qquad \overset{①}{\leq} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)} - \boldsymbol{w}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)} - \mathbf{g}^{(0)}\right\|^2 \\
&\qquad \overset{②}{=} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)} - \boldsymbol{w}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}^{(0)}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}_j^{(*)}\right\|^2\right]+8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2 \\
&\overset{③}{\leq} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[(1-2\mu\eta(1-\eta L))^c\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{\eta\sigma^2}{\mu(1-\eta L)}\right]+8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2 \\
&\overset{④}{\leq} \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\left[\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{\eta\sigma^2}{\mu(1-\eta L)}\right]+8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2 \\
&\overset{⑤}{\leq} \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{8L^4\eta^3\tau^2\sigma^2}{\mu(1-\eta L)}+16L^3\tau^2\eta^2\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right) \qquad (85)
\end{aligned}
$$

where ① comes from Assumption 1, ② holds because at the optimal local solution $\boldsymbol{w}_j^*$ of device $j$ we have $\mathbf{g}_j^{(*)}=\mathbf{0}$, ③ comes from strong convexity assumption for local cost functions where $\left\|\boldsymbol{w}_j^{(t,0)}-\boldsymbol{w}_j^{(*)}\right\|^2 \leq (1-2\mu\eta(1-\eta L))^t\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{\eta\sigma^2}{\mu(1-\eta L)}$ [34], ④ holds due to the choice of learning rate $\eta$ such that $(1-2\mu\eta(1-\eta L))^c \leq 1$, and finally ⑤ is due to smoothness assumption which implies $\left\|\mathbf{g}^{(0)}\right\|^2 \leq 2L\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)$ holds at global optimal solution $\boldsymbol{w}^{(*)}$. □

**Corollary E.8** (Linear speed up)**.** *To achieve linear speed up we set $\eta=\frac{1}{2L(q+1)\tau\gamma}$ and $\gamma\geq\sqrt{m\tau}$ in Eq. (78) which incurs:*

$$
\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big] \leq{}& \left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)e^{-\left(\frac{R}{6(q+1)\kappa}\right)} \\
&+\frac{3}{\mu}\Big[\frac{4.5\sigma^2}{\tau\gamma^2(q+1)^2}+\frac{2L^2}{(q+1)^2\gamma^2 m}\sum_{j=1}^{m}\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{2L\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)}{(q+1)^2\gamma^2} \\
&+\frac{\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau-0.5\right)}+\frac{5}{16(q+1)^3\tau\gamma^2}\frac{\sigma^2}{m}+\frac{5}{16(q+1)^3\tau^2\gamma^2}G_q \\
&+\frac{1}{4\tau}\frac{\sigma^2}{m}+\frac{G_q}{4(q+1)\tau^2}\Big] \qquad (86)
\end{aligned}
$$

*From Eq. (86) we can see that to attain an $\epsilon$-accurate solution we can choose*

$$
R=O\left(\kappa(q+1)\log\left(\frac{1}{\epsilon}\right)\right),\tau=O\left(\frac{1}{m\epsilon}\right),
$$

*as desired.*

## E.3 Main result for the general convex setting

**Theorem E.9** (Convex)**.** *For a convex function $f(\boldsymbol{w})$, applying* `FedCOMGATE`$(\tau,\eta,\gamma)$ *(Algorithm 2) to optimize $\tilde{f}(\boldsymbol{w},\phi)=f(\boldsymbol{w})+\frac{\phi}{2}\left\|\boldsymbol{w}\right\|^2$, for all $0\leq t\leq R\tau-1$, under Assumptions 1,2, 4, 5 if the learning rate satisfies*

$$1\geq 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4+(q+1)L\eta\gamma\tau \quad \& \quad 30\eta^2L^2\tau^2\leq 1 \tag{87}$$

*and all the models are initialized with $\boldsymbol{w}^{(0)}$, with the choice of $\phi=\frac{1}{\sqrt{m\tau}}$ and $\eta=\frac{1}{2L\gamma\tau(1+q)}$ we obtain:*

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big]\\
&\leq e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)\\
&\quad+\Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}}+\frac{6\sqrt{m\tau}L^2}{m(q+1)^2\gamma^2}\sum_{j=1}^{m}\left\|\Big(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\Big)\right\|_2^2+\frac{12L\sqrt{m\tau}}{\gamma^2(q+1)^2}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big).\\
&\quad+\frac{3\sqrt{m\tau}\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau-0.5\right)}+\frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}}+\frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2}+\frac{3\sigma^2}{4\sqrt{m\tau}}+\frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big]\\
&\quad+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned}\tag{88}$$

*Proof.* Since $\tilde{f}(\boldsymbol{w}^{(r)},\phi)=f(\boldsymbol{w}^{(r)})+\frac{\phi}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2$ is $\phi$-PL, according to Theorem E.6, we have:

$$\begin{aligned}
\tilde{f}(\boldsymbol{w}^{(R)},\phi)-\tilde{f}(\boldsymbol{w}^{(*)},\phi)&=f(\boldsymbol{w}^{(r)})+\frac{\lambda}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2-\left(f(\boldsymbol{w}^{(*)})+\frac{\lambda}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2\right)\\
&\leq\left(1-\frac{\eta\gamma\phi\tau}{3}\right)^R\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)\\
&\quad+\frac{3}{\phi}\Big[L^2 18\eta^2\tau\sigma^2+L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\right\|^2\\
&\quad+5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}+5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q+\frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m}+\frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned}\tag{89}$$

Next rearranging Eq. (89) and replacing $\mu$ with $\phi$, and using the short hand notation of

$$\begin{aligned}
\mathcal{A}(\eta)\triangleq\Big[&L^2 18\eta^2\tau\sigma^2+L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\right\|^2\\
&+5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}+5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q+\frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m}+\frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned}\tag{90}$$

leads to the following error bound:

$$\begin{aligned}
\tilde{f}(\boldsymbol{w}^{(R)},\phi)-f^*&\leq\left(1-\frac{\eta\gamma\phi\tau}{3}\right)^R\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)+\frac{3}{\phi}\mathcal{A}(\eta)+\frac{\phi}{2}\left(\left\|\boldsymbol{w}^{(*)}\right\|^2-\left\|\boldsymbol{w}^{(r)}\right\|^2\right)\\
&\leq e^{-\left(\frac{\eta\gamma\phi\tau}{3}\right)R}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)+\frac{3}{\phi}\mathcal{A}(\eta)+\frac{\phi}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned}\tag{91}$$

Next, if we set $\phi=\frac{1}{\sqrt{m\tau}}$ and $\eta=\frac{1}{2(1+q)L\gamma\tau}$, we obtain the following bound:

$$\tilde{f}(\boldsymbol{w}^{(R)},\phi)-f^*\leq e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)+3\sqrt{m\tau}\mathcal{A}\Big(\frac{1}{2\left(1+q\right)L\gamma\tau}\Big)+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2$$

$$
\begin{aligned}
&= e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}} \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}} + \frac{3}{\sqrt{m}\,(q+1)^2\,\gamma^2\tau^{2.5}} \sum_{j=1}^{m}\sum_{c=0}^{\tau-1} \left\| \sum_{c=0,r=0}^{\tau-1} \left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right) \right\|^2 \\
&\quad + \frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}} + \frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2} + \frac{3\sigma^2}{4\sqrt{m\tau}} + \frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big] \\
&\quad + \frac{1}{2\sqrt{m\tau}} \left\|\boldsymbol{w}^{(*)}\right\|^2 
\end{aligned} \tag{92}
$$

Finally, using Eq. (85) we obtain the bound:

$$
\begin{aligned}
\tilde{f}(\boldsymbol{w}^{(R)}, \phi) - f^* &\leq e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}} \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}} + \frac{6\sqrt{m\tau}L^2}{(q+1)^2\gamma^2 m} \sum_{j=1}^{m} \left\|\left(\boldsymbol{w}_j^{(0,0)} - \boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + \frac{12L\sqrt{m\tau}}{\gamma^2(q+1)^2} \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{3\sqrt{m\tau}\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau - 0.5\right)} + \frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}} + \frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2} + \frac{3\sigma^2}{4\sqrt{m\tau}} + \frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big] \\
&\quad + \frac{1}{2\sqrt{m\tau}} \left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned} \tag{93}
$$

$\square$

**Corollary E.10.** *As a result of Theorem E.9, for general convex functions with $\gamma \geq \sqrt{m\tau}$, to achieve the convergence error of $\epsilon$ we need to have $\tau = O\left(\frac{1}{m\epsilon^2}\right)$ and $R = O\left(\frac{L(1+q)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$.*

# F Deferred Proofs

## F.1 Proof of Lemma E.3

We prove Lemma E.3 in two steps. First, we prove the following lemma:

**Lemma F.1.** *Under Assumption 1 and 4, and the condition over learning rate $30\eta^2\tau^2L^2 \leq 1$, we have the following inequality:*

$$\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\left(\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(c,r)}\right)\right\|^2 = \frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\sum_{c=0,r}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2$$
$$\leq 36R\eta^2\tau^2\sigma^2 + 8\eta^2 C + 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2$$
$$+\frac{10\eta^2L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2 \tag{94}$$

*where* $C = \frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r=0}^{\tau-1}\left\|\sum_{k=0,r=0}^{c}\left(\nabla f_j(\boldsymbol{w}_j^{(k,r)}) - \nabla f(\boldsymbol{w}^{(r)})\right)\right\|^2$

First, we bound the term $\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\left(\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(c,r)}\right)\right\|^2$ for $r \geq 1$:

**Lemma F.2.** *For $r \geq 1$:*

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\tilde{d}_j^{(c,r)}\right\|^2 \leq 18\sigma^2\tau + \frac{1}{p}\sum_{j=1}^{p}\Big[6L^2\tau\Big[\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2 + \frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$
$$+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2 + \sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2$$
$$+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2 + \frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\Big] \tag{95}$$

*Proof.*

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\tilde{d}_j^{(c,r)}\right\|^2$$
$$=\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)}+\frac{1}{\tau}\left(\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}-\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\right]\right\|^2$$
$$=\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}+\mathbf{g}_j^{(c,r)}\right)$$
$$+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}+\mathbf{g}_j^{(c,r-1)}\right)-\mathbf{g}_j^{(c,r-1)}+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2$$
$$\leq\frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\underbrace{\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2}_{\text{(I)}}$$

$$+\frac{2}{p}\sum_{j=1}^{p}\underbrace{\Big\|\sum_{c=0}^{\tau-1}\Big[\mathbf{g}_j^{(c,r)}+\frac{2}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big]\Big\|^2}_{\text{(II)}} \tag{96}$$

□

We first bound the term (I) in Eq. (96) with the following lemma:

**Lemma F.3.**

$$\frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2\leq 18\sigma^2\tau \tag{97}$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2\\
&\leq \mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)\Big\|^2+\Big\|\frac{1}{\tau}\sum_{c=0}^{\tau-1}\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big\|^2+\Big\|\frac{1}{\tau}\sum_{c=0}^{\tau-1}\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big\|^2\\
&=3\left[\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big\|^2\right]\\
&\overset{①}{=}3\left[\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\left(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right)\Big\|^2+\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\left(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big\|^2+\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\left(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big\|^2\right]\\
&\leq\tau\left(\sigma^2+\sigma^2+\sigma^2\right)\\
&=9\sigma^2\tau
\end{aligned}\tag{98}$$

where ① follows from Assumption 4. □

We bound the term (II) in Eq. (96) as follows:

**Lemma F.4.** *For $r\geq 1$ we have:*

$$\Big\|\sum_{c=0,r}^{\tau-1}\Big[\mathbf{g}_j^{(c,r)}+\frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big]\Big\|^2 \tag{99}$$

$$\begin{aligned}
&\leq 5L^2\Big[\tau\sum_{c=0,r}^{\tau-1}\Big\|\Big[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\Big]\Big\|^2+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\Big\|^2\\
&\quad+\tau\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\tau\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2\\
&\quad+\tau\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\tau\frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^2\Big]
\end{aligned}\tag{100}$$

*Proof.* Adopting the notation $\mathbf{g}_j^{(r)} = \nabla f_j(\boldsymbol{w}^{(r)})$, we have:

$$
\begin{aligned}
&\Big\| \sum_{c=0}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \left( \frac{1}{p} \sum_{j=1}^{p} \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(c,r-1)} \right) \Big] \Big\|^2 \\
&= \Big\| \sum_{c=0}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} + \mathbf{g}_j^{(r)} \\
&\quad + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \left( -\mathbf{g}_j^{(r-1)} + \mathbf{g}_j^{(r-1)} - \mathbf{g}_j^{(c,r-1)} \right) + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} + \mathbf{g}_j^{(r-1)} \right) \Big] \Big\|^2 \\
&= \Big\| \sum_{c=0}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} + \mathbf{g}_j^{(r)} - \frac{1}{\tau} \sum_{c=0}^{\tau-1} \mathbf{g}_j^{(r-1)} \\
&\quad + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \left( \mathbf{g}_j^{(r-1)} - \mathbf{g}_j^{(c,r-1)} \right) + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) + \mathbf{g}^{(r-1)} \Big] \Big\|^2 \\
&\leq 5 \Big[ \Big\| \sum_{c=0,r}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \left( \mathbf{g}_j^{(r)} - \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(r-1)} - \mathbf{g}_j^{(c,r-1)} \right) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big] \\
&\overset{①}{=} 5 \Big[ \Big\| \sum_{c=0,r}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \Big] \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big]
\end{aligned} \tag{101}
$$

where ① holds due to $\frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \left( \mathbf{g}_j^{(c,r)} - \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \mathbf{g}_j^{(r)} \right) = \sum_{c=0,r}^{\tau-1} \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)} \right)$. We continue with bounding Eq. (101):

$$
\begin{aligned}
&5 \Big[ \Big\| \sum_{c=0,r}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \Big] \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left( \mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big] \\
&\overset{②}{=} 5 \Big[ \Big\| \sum_{c=0,r}^{\tau-1} \Big[ \mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \Big] \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \left( \mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\Big\| \sum_{c=0,r-1}^{\tau-1} \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 + \Big\| \sum_{c=0,r-1}^{\tau-1} \frac{1}{p}\sum_{j=1}^{p} \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 \\
&+\Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big] \\
\le\; & 5\Big[\tau \sum_{c=0,r}^{\tau-1} \Big\| \Big[\mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)}\Big] \Big\|^2 + \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \Big\| \Big(\mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 \\
&+ \sum_{c=0,r-1}^{\tau-1} \Big\| \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 + \sum_{c=0,r-1}^{\tau-1} \frac{1}{p}\sum_{j=1}^{p} \Big\| \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 \\
&+ \tau \sum_{c=0,r-1}^{\tau-1} \Big\| \mathbf{g}^{(r-1)} \Big\|^2 \Big] \\
\le\; & 5L^2\Big[\tau \sum_{c=0,r}^{\tau-1} \Big\| \Big[\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big] \Big\|^2 + \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \Big\| \boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)} \Big\|^2 \\
&+ \tau \sum_{c=0,r-1}^{\tau-1} \Big\| \boldsymbol{w}_j^{(c,r-1)} - \boldsymbol{w}^{(r-1)} \Big\|^2 + \tau \sum_{c=0,r-1}^{\tau-1} \frac{1}{p}\sum_{j=1}^{p} \Big\| \boldsymbol{w}_j^{(c,r-1)} - \boldsymbol{w}^{(r-1)} \Big\|^2 + \frac{\tau}{L^2} \sum_{c=0,r-1}^{\tau-1} \Big\| \mathbf{g}^{(r-1)} \Big\|^2 \Big] \qquad (102)
\end{aligned}
$$

where ② is due to the fact that

$$
\begin{aligned}
&\Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p}\sum_{j=1}^{p} \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 \\
&= \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \Big\| \Big(\mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2 + \Big\| \sum_{c=0,r-1}^{\tau-1} \frac{1}{p}\sum_{j=1}^{p} \Big(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)}\Big) \Big\|^2, \qquad (103)
\end{aligned}
$$

as $\mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)}$ depends on argument in round $r-1$. □

**Lemma F.5.** *For $r = 0$, we have:*

$$
\frac{1}{p}\sum_{j=1}^{p} \mathbb{E} \Big\| \sum_{c=0}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r)} \Big\|^2 \le \frac{4}{p}\sum_{j=1}^{p} \Big[ \tau\sigma^2 + \tau L^2 \sum_{c=0}^{\tau-1} \Big\| \boldsymbol{w}_j^{(c,0)} - \boldsymbol{w}^{(c,0)} \Big\|^2 + \Big\| \sum_{c=0}^{\tau-1} \Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(c,0)}\Big) \Big\|^2 + \tau \sum_{c=0}^{\tau-1} \Big\| \mathbf{g}^{(0)} \Big\|^2 \Big] \qquad (104)
$$

*Proof.* For $r = 0$ we have:

$$
\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p} \mathbb{E} \Big\| \sum_{c=0}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,0)} \Big\|^2 \\
&= \frac{1}{p}\sum_{j=1}^{p} \mathbb{E} \Big\| \sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,0)} \Big\|^2 \\
&= \frac{1}{p}\sum_{j=1}^{p} \mathbb{E} \Big\| \sum_{c=0}^{\tau-1} \Big(\tilde{\mathbf{g}}_j^{(c,0)} - \mathbf{g}_j^{(c,0)} + \mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(0)} + \mathbf{g}_j^{(0)} - \mathbf{g}^{(0)} + \mathbf{g}^{(0)}\Big) \Big\|^2 \\
&\le \frac{4}{p}\sum_{j=1}^{p} \Big[ \mathbb{E} \Big\| \sum_{c=0}^{\tau-1} \Big(\tilde{\mathbf{g}}_j^{(c,0)} - \mathbf{g}_j^{(c,0)}\Big) \Big\|^2 + \Big\| \sum_{c=0}^{\tau-1} \Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}_j^{(0)}\Big) \Big\|^2 + \Big\| \sum_{c=0}^{\tau-1} \Big(\mathbf{g}_j^{(0)} - \mathbf{g}^{(0)}\Big) \Big\|^2 + \Big\| \sum_{c=0}^{\tau-1} \mathbf{g}^{(0)} \Big\|^2 \Big]
\end{aligned}
$$

$$\overset{①}{=} \frac{4}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbb{E}\left\|\left(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}\right)\right\|^2+\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\right)\right\|^2+\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\right)\right\|^2+\left\|\sum_{c=0}^{\tau-1}\mathbf{g}^{(0)}\right\|^2\right]$$

$$\leq \frac{4}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbb{E}\left\|\left(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}\right)\right\|^2+\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\right)\right\|^2+\tau\sum_{c=0}^{\tau-1}\left\|\left(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\right)\right\|^2+\tau\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\right]$$

$$\leq \frac{4}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\sigma^2+\tau\sum_{c=0}^{\tau-1}\left\|\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\right)\right\|^2+\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\right)\right\|^2+\tau\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\right]$$

$$= \frac{4}{p}\sum_{j=1}^{p}\left[\tau\sigma^2+\tau L^2\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2+\tau\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\right] \tag{105}$$

where ① comes from i.i.d. mini-batch sampling. □

The rest of the proof comes from plugging both Lemmas F.4 and F.5 in Eq. (96) as shown below.

*Proof.*

$$\boldsymbol{w}_j^{(c,r)}=\boldsymbol{w}_j^{(c-1,r)}-\eta\tilde{\boldsymbol{d}}_j^{(c,r)}=\ldots=\boldsymbol{w}^{(r)}-\eta\sum_{k=0}^{c-1}\tilde{\boldsymbol{d}}_j^{(c,r)} \tag{106}$$

Now we can write:

$$\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right\|^2$$

$$=\frac{\eta^2}{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{k=0}^{c-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2$$

$$=\eta^2\left[\sum_{c=0,r=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{k=0}^{c-1}\tilde{\boldsymbol{d}}_j^{(c,0)}\right\|^2+\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{k=0}^{c-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2\right]$$

$$\leq\eta^2\Big(\sum_{c=0,r=0}^{\tau-1}\frac{4}{p}\sum_{j=1}^{p}\left[\tau\sigma^2+\tau L^2\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\left\|\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(c,0)}\right\|^2+\tau\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\right]$$

$$+\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\Big[18\sigma^2\tau+\frac{1}{p}\sum_{j=1}^{p}\Big[5L^2\Big[\tau\sum_{c=0,r}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right\|^2+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\tau\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2+\tau\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{\tau}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\Big]\Big)$$

$$=\eta^2\Big(\Big[\sum_{c=0,r=0}^{\tau-1}\frac{4\tau}{p}\sum_{j=1}^{p}\sigma^2+L^2\sum_{c=0,r=0}^{\tau-1}\frac{4\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2$$

$$+\sum_{c=0,r=0}^{\tau-1}\frac{4\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\Big]$$

$$+\Big[18\sigma^2\tau\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}1+\Big[5L^2\Big[\frac{\tau}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2+\frac{1}{\tau}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{\tau}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2+\frac{\tau}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\tau^2\sum_{r=1}^{R-1}\frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\Big]\Big)$$

$$=\eta^2\Big(\Big[4\tau^2\sigma^2+L^2\frac{4\tau^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2+4\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\Big]$$

$$+\Big[18\sigma^2(R-1)\tau^2+\Big[5L^2\Big[\frac{\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2+\tau\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2+\frac{\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\tau^2\sum_{r=1}^{R-1}\frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\Big]\Big) \tag{107}$$

Now we continue with bounding Eq. (107) with further simplification as follows:

$$=\eta^2\Big(\Big[4\tau^2\sigma^2+L^2\frac{4\tau^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2+4\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\Big]$$

$$+18\sigma^2(R-1)\tau^2+\frac{5L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right\|^2+5L^2\tau\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2+5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big)$$

$$\overset{①}{\leq}\eta^2\Big(\Big[18R\tau^2\sigma^2+L^2\frac{4\tau^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\right\|^2+\frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2+4\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\Big]$$

$$+\frac{5L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2+\frac{5L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2+5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big)$$

$$\overset{②}{\leq}\eta^2\Big(\Big[18R\tau^2\sigma^2+\frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2+5\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\Big]$$

$$+\frac{5L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2+\frac{5L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2$$

$$+\frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right\|^2+5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big)$$

$$
\overset{③}{\leq} 18R\eta^2\tau^2\sigma^2 + \frac{4\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right\|^2 + \frac{5\eta^2L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2
$$

$$
+ \frac{15\eta^2L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 + 10\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \tag{108}
$$

where ① comes from $4\tau^2\sigma^2 \leq 18\tau^2\sigma^2$, ② holds because of $4\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2 \leq 5\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2$ and ③ is due to

$$
5\eta^2\tau^2\sum_{c=0,r=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2 + 5\eta^2\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2 \leq 10\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2.
$$

Rearranging Eq. (108) we obtain:

$$
\begin{aligned}
&\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 \\
&\quad\leq \frac{18R\eta^2\tau^2\sigma^2}{1-15\eta^2L^2\tau^2} + \frac{4\eta^2}{p\left(1-15\eta^2L^2\tau^2\right)}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right\|^2 \\
&\quad+ \frac{5\eta^2L^2\tau}{p\left(1-15\eta^2L^2\tau^2\right)}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2 \\
&\quad\quad+ \frac{10\eta^2\tau^2}{\left(1-15\eta^2L^2\tau^2\right)}\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \\
&\quad\overset{①}{\leq} 36R\eta^2\tau^2\sigma^2 + \frac{36\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\quad+ \frac{10\eta^2L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2 \\
&\quad\quad+ 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2
\end{aligned} \tag{109}
$$

where ① comes from the condition $1 \geq 30\eta^2.L^2\tau^2$ ☐

**Lemma F.6.** *Under Assumptions 1, 2, 4 and 5 we have:*

$$
\begin{aligned}
\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\mathbb{E}_{\xi}\mathbb{E}_{Q}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2 &\leq \tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\right\|^2\right] \\
&+ \tau^3R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p} + \tau^2(\eta\gamma)^2(q+1)RG_q
\end{aligned} \tag{110}
$$

*Proof.*

$$
\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\mathbb{E}_{\xi}\mathbb{E}_{Q}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2
$$

$$
\begin{aligned}
&= \sum_{r=0}^{R-1} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \mathbb{E}_Q \left\| \boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)} \right\|^2 \\
&= \tau \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \mathbb{E}_Q \left\| \boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)} \right\|^2 \\
&= \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \mathbb{E}_Q \Big\| \frac{1}{p} \sum_{j=1}^{p} Q \left( \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) \Big\|^2 \\
&\overset{①}{\leq} \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \left[ \mathbb{E}_Q \Big\| Q \left( \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) \Big\|^2 + G_q \right] \\
&\overset{②}{\leq} \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \Big[ \mathbb{E}_Q \Big\| Q \left( \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) - \mathbb{E}_Q \left[ Q \left( \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) \right] \Big\|^2 \\
&\quad + \left\| \mathbb{E}_Q \left[ Q \left( \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) \right] \right\|^2 + G_q \Big] \\
&= \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \Big[ \mathbb{E}_Q \Big\| Q \left( \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right) - \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \Big\|^2 \\
&\quad + \left\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \right\|^2 + G_q \Big] \\
&\overset{③}{\leq} \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} \mathbb{E}_\xi \Big[ q \Big\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \Big\|^2 + \left\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \right\|^2 + G_q \Big] \\
&= \tau(\eta\gamma)^2 \sum_{r=1}^{R-1} \sum_{c=0,r-1}^{\tau-1} (q+1) \mathbb{E}_\xi \Big[ \Big\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \Big\|^2 + G_q \Big] \\
&= \tau^2(\eta\gamma)^2 (q+1) \sum_{r=1}^{R-1} \mathbb{E}_\xi \Big[ \Big\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r-1)} \right] \Big\|^2 + G_q \Big] \\
&\overset{④}{=} \tau^2(\eta\gamma)^2 (q+1) \sum_{r=1}^{R-1} \mathbb{E}_\xi \Big[ \Big\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r-1)} \right] \Big\|^2 + G_q \Big]
\end{aligned} \tag{111}
$$

where ① comes from Assumption 5, ② is due to the definition of variance, ③ holds because of Assumption 2 and ④ is because of $\frac{1}{p}\sum_{j=1}^{p} \delta^{(r,\tau)} = 0$. We continue from Eq. (111) as follows:

$$
\begin{aligned}
&= \tau^2(\eta\gamma)^2 (q+1) \sum_{r=1}^{R-1} \mathbb{E}_\xi \Big[ \Big\| \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r-1)} \right] \Big\|^2 \Big] + \tau^2(\eta\gamma)^2 (q+1) R G_q \\
&= \tau^2(\eta\gamma)^2 (q+1) \sum_{r=1}^{R-1} \mathrm{Var}_\xi \Big( \left[ \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r-1)} \right] \Big) + \tau^2(\eta\gamma)^2 (q+1) \sum_{r=1}^{R-1} \left\| \mathbb{E}_\xi \frac{1}{p} \sum_{j=1}^{p} \sum_{c=0,r-1}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r-1)} \right\|^2 \\
&\quad + \tau^2(\eta\gamma)^2 (q+1) R G_q
\end{aligned}
$$

$$
\begin{aligned}
&= \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathrm{Var}_\xi\Big(\Bigg[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Bigg]\Big) + \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\Bigg\|\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathbf{g}_j^{(c,r-1)}\Bigg\|^2 \\
&\qquad + \tau^2(\eta\gamma)^2(q+1)RG_q \\
&\overset{①}{=} \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathrm{Var}_\xi\Big(\Big[\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big) + \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\Bigg\|\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathbf{g}_j^{(c,r-1)}\Bigg\|^2 \\
&\qquad + \tau^2(\eta\gamma)^2(q+1)RG_q \\
&\overset{②}{\leq} \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\sigma^2 + \tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}_j^{(c,r-1)}\Big\|^2 + \tau^2(\eta\gamma)^2(q+1)RG_q \\
&= \tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}_j^{(c,r-1)}\Big\|^2 + \tau^3(\eta\gamma)^2(q+1)R\frac{1}{p}\sigma^2 + \tau^2(\eta\gamma)^2(q+1)RG_q
\end{aligned} \tag{112}
$$

where ① comes from i.i.d. mini-batch sampling and ② is due to inequlity $\|\sum_{i=1}^{n}\mathbf{a}_i\|^2 \leq n\sum_{i=1}^{n}\|\mathbf{a}_i\|^2$. $\square$

Finally, by plugging Lemma F.6 into Eq. (109), we obtain the following bound:

$$
\begin{aligned}
&\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big\|^2 \\
&\quad \leq 36R\eta^2\tau^2\sigma^2 + \frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Bigg\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Bigg\|^2 \\
&\qquad + 10\eta^2L^2\tau\Bigg[\tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big[\Big\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\Big\|^2\Big] + \tau^3R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p} + \tau^2(\eta\gamma)^2(q+1)RG_q\Bigg] \\
&\qquad + 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(r)}\Big\|^2 \\
&\quad = 36R\eta^2\tau^2\sigma^2 + \frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Bigg\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Bigg\|^2 \\
&\qquad + \Big[10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big[\Big\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\Big\|^2\Big] \\
&\qquad + 10\eta^2L^2\tau^4R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p} + 10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)RG_q\Big] \\
&\qquad + 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(r)}\Big\|^2
\end{aligned} \tag{113}
$$

### F.2 Proof of Lemma E.7

Similarly, using Lemmas F.2 and F.5 for every communication round we can write:

$$
\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\mathbb{E}\left\|\boldsymbol{w}_j^{(r,c)}-\boldsymbol{w}^{(r)}\right\|^2 \\
&\leq 36\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2 \\
&+10\eta^2L^2\tau\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2+20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \\
&\overset{①}{\leq} 36\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2 \\
&+\Bigg[10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\right\|^2\right] \\
&\quad+10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{p}+10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q\Bigg]+20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2
\end{aligned}
\tag{114}
$$

where ① follows from Lemma F.6 without summation over $r$.

# References


[1] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 440–445, 2017.

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[3] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

[4] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pages 14668–14679, 2019.

[5] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569, 2018.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[7] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[8] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

[9] Robert Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, 2019.

[10] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 2019.

[11] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Trading redundancy for communication: Speeding up distributed sgd for non-convex optimization. In *ICML*, pages 2545–2554, 2019.

[12] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[15] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. In *Advances in Neural Information Processing Systems*, pages 13144–13154, 2019.

[16] Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

[17] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[18] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

[20] A Khaled, K Mishchenko, and P Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

[21] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019.

[22] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

[23] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[24] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[25] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. *arXiv preprint arXiv:2002.11364*, 2020.

[26] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

[27] Sen Lin, Guang Yang, and Junshan Zhang. A collaborative learning framework via federated meta-learning. *arXiv preprint arXiv:2001.03229*, 2020.

[28] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

[29] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *ICLR*, 2018.

[30] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

[31] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

[32] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

[33] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[34] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[36] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[37] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031, 2020.

[38] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Quantized decentralized consensus optimization. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5838–5843. IEEE, 2018.

[39] Amirhossein Reisizadeh, Hossein Taheri, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Robust and communication-efficient collaborative learning. In *Advances in Neural Information Processing Systems*, pages 8388–8399, 2019.

[40] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. "The" MIT Press, 2002.

[41] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[42] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. Squarm-sgd: Communication-efficient momentum sgd for decentralized optimization. *arXiv preprint arXiv:2005.07041*, 2020.

[43] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017.

[44] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[45] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[46] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[47] Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3329–3337. JMLR. org, 2017.

[48] Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems*, pages 7652–7662, 2018.

[49] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

[50] Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. Matcha: Speeding up decentralized sgd via matching decomposition sampling. *arXiv preprint arXiv:1905.09435*, 2019.

[51] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309, 2018.

[52] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pages 1509–1519, 2017.

[53] Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020.

[54] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized sgd and its applications to large-scale distributed optimization. In *International Conference on Machine Learning*, pages 5325–5333, 2018.

[55] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd for non-convex optimization with faster convergence and less communication. *arXiv preprint arXiv:1807.06629*, 2018.

[56] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[57] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.

[58] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

[59] Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3219–3227, 2018.