Supplementary Material
Federated Learning with Compression: Unified Analysis and Sharp Guarantees

The outline of our supplementary material follows. In Section A, we first elaborate further on related studies in the literature. In Section B.1, 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 and Comparison

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 [69], also known as local SGD. The seminal work of [55] 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 [14,15,27,57,60,68]. In [60], 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 [14,38,51].

The analysis of local SGD in the heterogeneous setting, also known as federated averaging (FedAvg) [43], has seen a resurgence of interest very recently. While it is still an active research area [64], a few number of recent studies made efforts to understand the convergence of local SGD in a heterogeneous setting [11,18,26,27,30,33,34,71]. 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 [10,12,37,42,54].

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. [5,40,52,57], a quantization operator is applied before transmitting the gradient to server. A gradient acceleration approach with compression is proposed in [35]. In heterogeneous data distribution, [23] proposed the use of sign based SGD algorithms and [48] 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 [2,41]. For these methods, theoretical guarantees have been provided in a few recent efforts [4,19,28,56,59,65]. 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 [17,22].

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 [29,49,50,53]. Another interesting direction for the purpose of reducing the communication complexity is to exploit the sparsity of communication network as explored in [18,30,61]. Moreover, the recent work of [46] studies bidirectional compression in Federated Learning.

Finally, more thorough related works that study federated learning from different perspectives can be found in [24] and [32].

Additional comparison with DIANA [21] We provide a summary of the comparison of our algorithms and algorithms introduced in [21] in two tables. We compare the rates in homogeneous and heterogeneous data distributions separately. The following comments are in place:

In the homogeneous setting, shown in Table 5 and in comparison to DIANA and VR-DIANA, FedCOM improves all the communication rounds in terms of dependency on q (shown in blue). In the heterogeneous setting, shown in Table 6, in comparison to DIANA and VR-DIANA, FedCOMGATE basically improves all the communication rounds in terms of dependency on q (shown in blue) except for the strongly convex (SC) case. For the SC case of heterogeneous setting, we highlight that our results are for the PL, unlike DIANA which is for SC. Thus, we believe that if we derive the results directly for SC we might obtain the same or even better results than DIANA (like homogeneous setting). Comparison of finite-sum and stochastic algorithms does not seem to be fair, but per your request, we provide the full comparison. We believe if we analyze our methods for FS settings the dependency on n would only appear in τ (not R). For deterministic settings, i.e., setting $n = 1$ in DIANA and $\sigma^2 = 0$ in FedCOM and FedCOMGATE, again we observe that the communication bounds for FedCOM and FedCOMGATE are better in terms of dependency on q in homogeneous and heterogeneous settings except for the SC heterogeneous case.

Reference	Objective function				F.S.
	Nonconvex	PL	Strongly Convex	General Convex	
DIANA [21]	–	–	$R = \tilde{O}\left(\kappa + \frac{\kappa q}{m} + q\right)$ $\tau = 1$	–	✗
VR-DIANA [21]	$R = O\left(\frac{\left(1+\frac{q}{m}\right)^{\frac{1}{2}}\left(n^{2/3}+q\right)}{\epsilon}\right)$ $\tau = 1$	–	$R = \tilde{O}\left(\kappa + \frac{\kappa q}{m} + q + n\right)$ $\tau = 1$	$R = O\left(\frac{\left(1+\frac{q}{m}\right)\sqrt{n}+\frac{q}{\sqrt{n}}}{\epsilon}\right)$ $\tau = 1$	✔
FedCOM [16]	$\boldsymbol{R = O\left(\frac{1}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{\frac{q}{m}+1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\kappa + \frac{\kappa q}{m}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\kappa + \frac{\kappa q}{m}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\frac{1+\frac{q}{m}}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon^2}\right)}$	✗

Table 5: **Homogeneous** data distribution with R communication rounds and τ local updates. F.S. stands for finite-sum assumption, $n = \max_{i\in[m]} n_i$, where n_i is the number of local samples at the ith device. m is the total number of devices, and q is the quantization noise. We use $\tilde{O}(.)$ to keep key parameters and to omit $\log(\frac{1}{\epsilon})$ term.

Reference	Objective function				F.S.
	Nonconvex	PL	Strongly Convex	General Convex	
DIANA [21]	–	–	$R = \tilde{O}\left(\kappa + \frac{\kappa q}{m} + q\right)$ $\tau = 1$	–	✗
VR-DIANA [21]	$R = O\left(\frac{\left(1+\frac{q}{m}\right)^{\frac{1}{2}}\left(n^{2/3}+q\right)}{\epsilon}\right)$ $\tau = 1$	–	$R = \tilde{O}\left(\kappa + \frac{\kappa q}{m} + q + n\right)$ $\tau = 1$	$R = O\left(\frac{\left(1+\frac{q}{m}\right)\sqrt{n}+\frac{q}{\sqrt{n}}}{\epsilon}\right)$ $\tau = 1$	✔
FedCOMGATE [16]	$\boldsymbol{R = O\left(\frac{q+1}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\kappa\left(q+1\right)\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\kappa\left(q+1\right)\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon}\right)}$	$\boldsymbol{R = \tilde{O}\left(\frac{1+q}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{m\epsilon^2}\right)}$	✗

Table 6: **Heterogeneous** data distribution with R communication rounds and τ local updates. F.S. stands for finite-sum assumption, $n = \max_{i\in[m]} n_i$, where n_i is the number of local samples at the ith device. m is the total number of devices, and q is the quantization noise. We use $\tilde{O}(.)$ to keep key parameters and to omit $\log(\frac{1}{\epsilon})$ term. Note that our results for PL condition hold for the strongly convex case as the latter is implied by former.

B Further Experimental Studies and Results

In this section, we present additional experimental results, as well as details, that will further showcase the efficacy of the proposed algorithms in the paper. First, we should elaborate on the algorithms we used in Section 6, but due to lack of space we did not describe in the main body. In addition, we introduce a version of Algorithm 2 without compression, and its version with sampling of clients.

B.1 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 [26], 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× speedup in the communication. Note that, since the communication time of broadcasting from server 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 [48].

Algorithm 3: `FedGATE`(R, τ, η, γ) Federated Averaging with Local Gradient Tracking

Inputs: Number of communication rounds R, number of local updates τ, learning rates γ and η, 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

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.

B.2 Additional Experiments

EMNIST dataset In addition to the results presented for the datasets in the main body, here, we present the results of applying different algorithms on EMNIST [9] dataset. This dataset, similar to the MNIST dataset, contains images of characters in 28×28 size. The difference here is that the dataset is separated based on the author of images, hence, the distribution each image is coming from is different for different nodes. In this experiment, we use data from 1000 authors in the EMNIST dataset, and set the sampling ratio $k = 0.1$. Also, we tune the learning rate to the fixed value of 0.01 for all the algorithms. The model, similar to the MNIST case, is a 2-layer MLP with 200 neurons for each hidden layer and ReLU activations. Figure 6 shows the results of this experiment for the training loss and testing accuracy based on the size of communication. It can be inferred that `FedCOMGATE` and FedPAQ both have the fastest convergence based on the communication size, and accordingly, wall-clock time. The reason that the final convergence rate is the same for all algorithms is that

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

Inputs: Number of communication rounds R, number of local updates τ, learning rates γ and η, 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

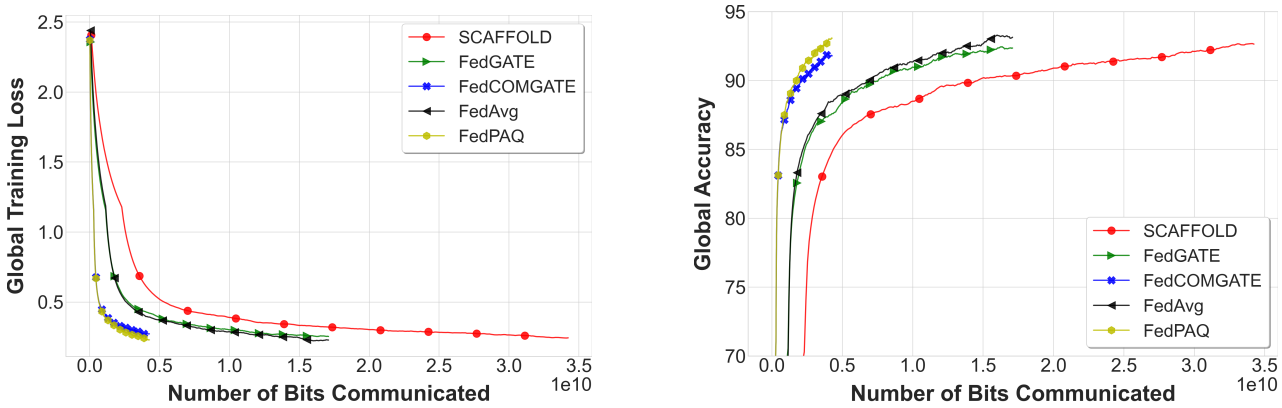
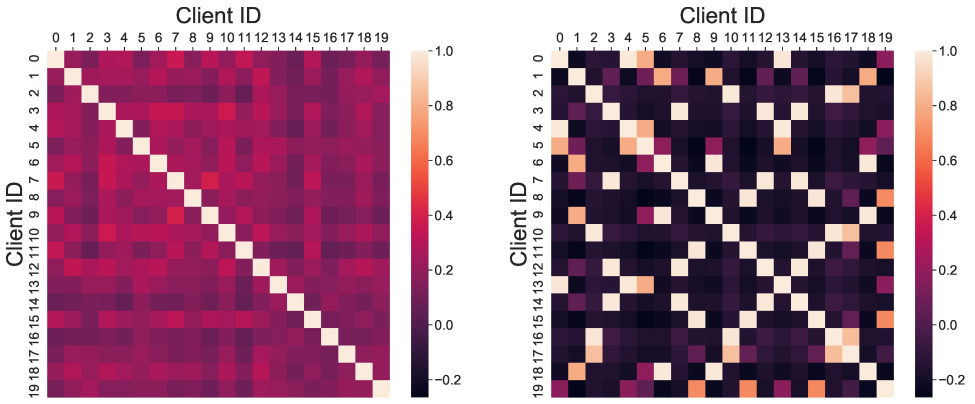


Figure 6: Comparing the performance of different algorithms on the EMNIST dataset, using 1000 clients' data on a 2-layer MLP model. `FedCOMGATE` and FedPAQ have the fastest convergence in time.

similar to Figure 2(a), this dataset is close to the homogeneous setting. To show that, and to compare it to the heterogeneous datatset we created using the MNIST dataset, we run a test on 20 different clients of this dataset and the heterogeneous MNIST dataset (2 classes data per client). We give all the clients the same model and perform a full batch gradient computation over that model. Then, we compute the cosine similarity of this gradients using:

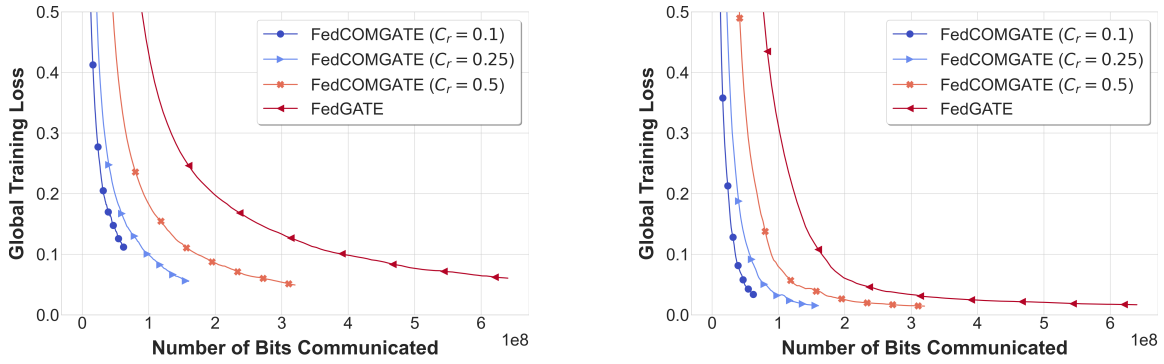
$$d_{ij} = \frac{\mathbf{g}_i^\top \mathbf{g}_j}{\|\mathbf{g}_i\| \cdot \|\mathbf{g}_j\|}. \tag{7}$$

Figure 7 shows the heatmap of these correlations among clients for two datasets. As it can be seen, in the EMNIST dataset, each client's data homogeneously correlates with all other clients'. However, in the MNIST dataset (with 2 classes data per client), each client has high correlation with at most 4 clients and not correlated or has a negative correlation with other clients' data. This shows that the level of heterogeneity in the EMNIST dataset is much lower than that of in the MNIST dataset, and hence, the result in the Figure 6 are in line with our theoretical findings for the gradient tracking technique.



(a) EMNIST dataset (b) MNIST dataset

Figure 7: Cosine similarity between full-gradients of different clients on the same model on EMNIST dataset and the heterogeneous MNIST (with 2 classes per client) dataset. In the EMNIST dataset each client has a homogeneous correlation with other clients, while the MNIST dataset is highly heterogeneous.



(a) MNIST Dataset (b) Fashion MNIST Dataset

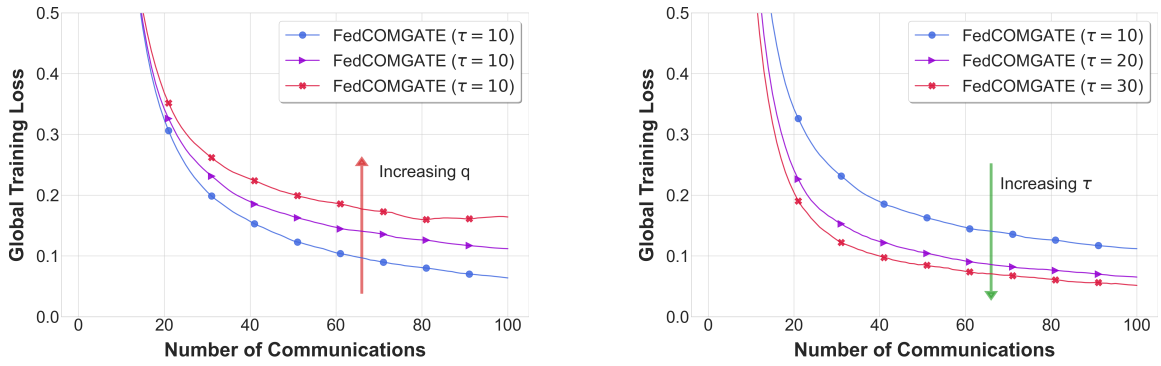
Figure 8: The effect of sparsification with memory on the `FedCOMGATE` algorithm used for the training of the `MNIST` and the `Fashion MNIST` datasets. We can achieve almost similar results as the algorithm without compression (`FedGATE`) with some compression rates. Decreasing the size of communication will speed up the training, in the cost of increasing a residual error as it is evident for the case with $C_r = 0.1$.

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 [2, 56, 62]. 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 [55] 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:

$$\boldsymbol{\Delta}_{j,s}^{(r)} = top_k\left\{\left(\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)}\right) + \boldsymbol{\nu}_j^{(r)}\right\} \tag{8}$$

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:

$$\boldsymbol{\nu}_j^{(r+1)} = \boldsymbol{\nu}_j^{(r)} + \frac{1}{m}\left(\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)}\right) - \boldsymbol{\Delta}^{(r)}, \tag{9}$$



(a) Effect of quantization noise q

(b) Effect of local computation τ

Figure 9: 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 τ to diminish the effects of quantization.

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 8 shows the result of this algorithm by changing the compression rate. As it was observed by [56], in some compression 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 9(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 9(b), where we keep the quantization noise constant and increase the number of local computations.

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 μ 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 (μ-strong convexity). *A function f is μ-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 ith data shard, respectively, where $\mathcal{Z}_i \subseteq \mathcal{S}_i$ is a uniformly sampled mini-bath. The corresponding quantities evaluated at ith machine's local solution at tth 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 cth local update at rth 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:*

$$\mathbb{E}_{Q,\xi^{(r)}}\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right] = \mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right] \leq \tau(q+1)\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + (q+1)\frac{\tau\sigma^2}{m} \tag{10}$$

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[\sum_{j=1}^{m}\frac{q}{m^2}\left\|\tilde{\mathbf{g}}_j^{(r)}\right\|^2 + \left\|\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_j^{(r)}\right\|^2\right] \\
&= \left[\sum_{j=1}^{m}\frac{q}{m^2}\left[\mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(r)}\right) + \left\|\mathbf{g}_j^{(r)}\right\|^2\right] + \left[\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{g}}_j^{(r)}\right) + \left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(r)}\right\|^2\right]\right] \\
&= \sum_{j=1}^{m}\frac{q}{m^2}\left[\mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(r)}\right) + \left\|\mathbf{g}_j^{(r)}\right\|^2\right] + \left[\frac{1}{m^2}\sum_{j=1}^{m}\mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(r)}\right) + \left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(r)}\right\|^2\right] \\
&\leq \sum_{j=1}^{m}\frac{q}{m^2}\left[\mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(r)}\right) + \left\|\mathbf{g}_j^{(r)}\right\|^2\right] + \left[\frac{1}{m^2}\sum_{j=1}^{m}\mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(r)}\right) + \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(r)}\right\|^2\right]
\end{aligned} \tag{11}$$

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

$$\mathbb{E}_{\xi^{(r)}}\left[\left[\|\tilde{\mathbf{g}}_j^{(r)} - \mathbf{g}_j^{(r)}\|^2\right]\right] \leq \tau\sigma^2 \tag{12}$$

To do so, note that

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

$$
\begin{aligned}
&= \mathrm{Var}\left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)}\right) \\
&\overset{②}{=} \sum_{c=0}^{\tau-1} \mathrm{Var}\left(\tilde{\mathbf{g}}_j^{(c,r)}\right) \\
&= \sum_{c=0}^{\tau-1} \mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)}\right\|^2\right] \\
&\overset{③}{\leq} \tau\sigma^2
\end{aligned} \tag{13}
$$

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 (11) by its upper bound in (12) implies that

$$
\mathbb{E}_{\xi^{(r)}|\boldsymbol{w}^{(r)}}\mathbb{E}_Q\Big[\|\frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\|^2\Big] \leq \sum_{j=1}^{m}\frac{q}{m^2}\left[\tau\sigma^2 + \left\|\mathbf{g}_j^{(r)}\right\|^2\right] + \left[\frac{1}{m^2}\sum_{j=1}^{m}\tau\sigma^2 + \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(r)}\right\|^2\right] \tag{14}
$$

Further note that we have

$$
\left\|\mathbf{g}_j^{(r)}\right\|^2 = \|\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\|^2 \leq \tau\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 \tag{15}
$$

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 (14) leads to the following bound:

$$
\mathbb{E}_{\xi^{(r)}|\boldsymbol{w}^{(r)}}\mathbb{E}_Q\Big[\|\frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\|^2\Big] \leq \tau(\frac{q}{m}+1)\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + (q+1)\frac{\tau\sigma^2}{m}, \tag{16}
$$

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:*

$$
-\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right] \leq \frac{1}{2}\eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2\|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2\right] \tag{17}
$$

Proof. We have:

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

$$\overset{②}{\leq} \frac{1}{2}\eta \sum_{c=0}^{\tau-1} \frac{1}{m} \sum_{j=1}^{m} \left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2 \|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2 \right] \tag{18}$$

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 rth communication round.

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

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

Proof. Note that

$$\begin{aligned}
\mathbb{E}\left[\left\| \boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)} \right\|_2^2 \right] &= \mathbb{E}\left[\left\| \boldsymbol{w}^{(r)} - \left(\boldsymbol{w}^{(r)} - \eta \sum_{k=0}^{c} \tilde{\mathbf{g}}_j^{(k,r)} \right) \right\|_2^2 \right] \\
&= \mathbb{E}\left[\left\| \eta \sum_{k=0}^{c} \tilde{\mathbf{g}}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{①}{=} \mathbb{E}\left[\left\| \eta \sum_{k=0}^{c} \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \left[\left\| \eta \sum_{k=0}^{c} \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{②}{=} \eta^2 \sum_{k=0}^{c} \mathbb{E}\left[\left\| \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + (c+1)\,\eta^2 \sum_{k=0}^{c} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\leq \eta^2 \sum_{k=0}^{\tau-1} \mathbb{E}\left[\left\| \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \tau\eta^2 \sum_{k=0}^{\tau-1} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\overset{③}{\leq} \eta^2 \sum_{k=0}^{\tau-1} \sigma^2 + \tau\eta^2 \sum_{k=0}^{\tau-1} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&= \eta^2 \tau \sigma^2 + \eta^2 \sum_{k=0}^{\tau-1} \tau \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2
\end{aligned} \tag{20}$$

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`(τ, η, γ)*, for all* $0 \leq t \leq R\tau - 1$*, under Assumptions 1 to 3, if the learning rate satisfies*

$$1 \geq \tau^2 L^2 \eta^2 + \left(\frac{q}{m} + 1 \right) \eta\gamma L\tau \tag{21}$$

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

$$\frac{1}{R} \sum_{r=0}^{R-1} \left\| \nabla f(\boldsymbol{w}^{(r)}) \right\|_2^2 \leq \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)}) \right)}{\eta\gamma\tau R} + \frac{L\eta\gamma(q+1)}{m}\sigma^2 + L^2\eta^2\tau\sigma^2 \tag{22}$$

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

$$\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(\tau,r)} = \eta \sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)}. \tag{23}$$

From the updating rule of Algorithm 1 we have

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \gamma\eta \left(\frac{1}{m} \sum_{j=1}^{m} Q\Big(\sum_{c=0,r}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \Big) \right) = \boldsymbol{w}^{(r)} - \gamma \left[\frac{\eta}{m} \sum_{j=1}^{m} Q \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right] \tag{24}$$

In what follows, we use the following notation to denote the stochastic gradient used to update the global model at rth 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:

$$\mathbb{E}_Q \left[\tilde{\mathbf{g}}_Q^{(r)} \right] = \frac{1}{m} \sum_{j=1} \left[-\eta \mathbb{E}_Q \left[Q \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right] \right] = \frac{1}{m} \sum_{j=1} \left[-\eta \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right] \triangleq \tilde{\mathbf{g}}^{(r)} \tag{25}$$

From the L-smoothness gradient assumption on global objective, by using $\tilde{\mathbf{g}}^{(r)}$ in inequality (23) we have:

$$f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)}) \leq -\gamma \langle \nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)} \rangle + \frac{\gamma^2 L}{2} \|\tilde{\mathbf{g}}^{(r)}\|^2 \tag{26}$$

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

$$\begin{aligned} \mathbb{E} \left[\mathbb{E}_Q \Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)}) \Big] \right] &\leq -\gamma \mathbb{E} \left[\mathbb{E}_Q \left[\langle \nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}_Q^{(r)} \rangle \right] \right] + \frac{\gamma^2 L}{2} \mathbb{E} \left[\mathbb{E}_Q \|\tilde{\mathbf{g}}_Q^{(r)}\|^2 \right] \\ &\overset{(a)}{=} -\gamma \underbrace{\mathbb{E} \left[\left[\langle \nabla f(\boldsymbol{w}^{(r)}), \tilde{\mathbf{g}}^{(r)} \rangle \right] \right]}_{\text{(I)}} + \frac{\gamma^2 L}{2} \underbrace{\mathbb{E} \left[\mathbb{E}_Q \Big[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2 \Big] \right]}_{\text{(II)}} \end{aligned} \tag{27}$$

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 (27), which gives

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

$$
\begin{aligned}
&= -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \\
&\quad - \left(1 - \tau L^2\eta^2\tau - (\frac{q}{m}+1)\eta\gamma L\tau\right)\frac{\eta\gamma}{2m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\|\mathbf{g}_j^{(c,r)}\|^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(q+1)\right)\sigma^2 \\
&\overset{②}{\leq} -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(q+1)\right)\sigma^2
\end{aligned} \tag{28}
$$

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

$$
1 \geq \tau L^2\eta^2\tau + (\frac{q}{m}+1)\eta\gamma L\tau. \tag{29}
$$

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

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

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. (22) for the choice of* $\eta\gamma = O\left(\frac{1}{L}\sqrt{\frac{m}{R\tau(q+1)}}\right)$, *and* $\gamma \geq m$ *the convergence rate reduces to:*

$$
\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq O\left(\frac{L\sqrt{(q+1)}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{*})\right)}{\sqrt{mR\tau}} + \frac{\left(\sqrt{(q+1)}\right)\sigma^2}{\sqrt{mR\tau}} + \frac{m\sigma^2}{R\gamma^2}\right). \tag{31}
$$

Note that according to Eq. (31), if we pick a fixed constant value for γ, *in order to achieve an* ϵ*-accurate solution,* $R = O\left(\frac{1}{\epsilon}\right)$ *communication rounds and* $\tau = O\left(\frac{q+1}{m\epsilon}\right)$ *local updates are necessary. We also highlight that Eq. (31) also allows us to choose* $R = O\left(\frac{q+1}{\epsilon}\right)$ *and* $\tau = O\left(\frac{1}{m\epsilon}\right)$ *to get the same convergence rate.*

Remark 7. *Condition in Eq. (21) can be rewritten as*

$$
\begin{aligned}
\eta &\leq \frac{-\gamma L\tau\left(\frac{q}{m}+1\right) + \sqrt{\gamma^2\left(L\tau\left(\frac{q}{m}+1\right)\right)^2 + 4L^2\tau^2}}{2L^2\tau^2} \\
&= \frac{-\gamma L\tau\left(\frac{q}{m}+1\right) + L\tau\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2 + 4}}{2L^2\tau^2} \\
&= \frac{\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2 + 4} - \left(\frac{q}{m}+1\right)\gamma}{2L\tau}
\end{aligned} \tag{32}
$$

So based on Eq. (32), if we set $\eta = O\left(\frac{1}{L\gamma}\sqrt{\frac{m}{R\tau(q+1)}}\right)$, *it implies that:*

$$
R \geq \frac{\tau m}{(q+1)\gamma^2\left(\sqrt{\left(\frac{q}{m}+1\right)^2\gamma^2 + 4} - \left(\frac{q}{m}+1\right)\gamma\right)^2} \tag{33}
$$

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

$$
R \geq \frac{\tau m}{5(q+1)} = O\left(\frac{\tau m}{q+1}\right) \tag{34}
$$

Therefore, for the choice of $\tau = O\left(\frac{q+1}{m\epsilon}\right)$, ***due to condition in Eq. (34), we need to have*** $R = O\left(\frac{1}{\epsilon}\right)$. ***Similarly, we can have*** $R = O\left(\frac{q+1}{\epsilon}\right)$ ***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. (22) reduces to*

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right)}{\eta R\tau} + \frac{L\eta}{m}\sigma^2 + L^2\eta^2\tau\sigma^2 \tag{35}$$

which matches the rate obtained in [60]. In this case the communication complexity and the number of local updates become

$$R = O\left(\frac{m}{\epsilon}\right), \quad \tau = O\left(\frac{1}{\epsilon}\right). \tag{36}$$

This simply implies that in this special case the convergence rate of our algorithm reduces to the rate obtained in [60], 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`(τ, η, γ), for all $0 \leq t \leq R\tau - 1$, under Assumptions 1 to 3 and 6,if the learning rate satisfies*

$$1 \geq \tau^2 L^2\eta^2 + \left(\frac{q}{m} + 1\right)\eta\gamma L\tau \tag{37}$$

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

$$\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq (1 - \eta\gamma\mu\tau)^R\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + (1+q)\frac{\gamma\eta L\sigma^2}{2m}\right] \tag{38}$$

Proof. From Eq. (28) under condition:

$$1 \geq \tau L^2\eta^2\tau + (\frac{q}{m} + 1)\eta\gamma L\tau \tag{39}$$

we obtain:

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\right] &\leq -\eta\gamma\frac{\tau}{2}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(q+1)\right)\sigma^2 \\ &\leq -\eta\mu\gamma\tau\left(f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(r)})\right) + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + \gamma(q+1)\right)\sigma^2\end{aligned} \tag{40}$$

which leads to the following bound:

$$\mathbb{E}\left[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(*)})\right] \leq (1 - \eta\mu\gamma\tau)\left[f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(*)})\right] + \frac{L\tau\gamma\eta^2}{2m}\left(mL\tau\eta + (\frac{q}{m} + 1)\gamma\right)\sigma^2 \tag{41}$$

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

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

□

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:*

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq e^{-\eta\gamma\mu\tau R}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}\tau L^2\eta^2\sigma^2 + (1+q)\,\frac{\gamma\eta L\sigma^2}{2m}\right] \\
&\leq e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{1}{\mu}\left[\frac{1}{2}L^2\frac{\tau\sigma^2}{L^2\left(\frac{q}{m}+1\right)^2\gamma^2\tau^2} + (1+q)\,\frac{L\sigma^2}{2\left(\frac{q}{m}+1\right)L\tau m}\right] \\
&= O\left(e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\sigma^2}{\left(\frac{q}{m}+1\right)^2\gamma^2\mu\tau} + \frac{(q+1)\,\sigma^2}{\mu\left(\frac{q}{m}+1\right)\tau m}\right) \\
&= O\left(e^{-\frac{R}{2\left(\frac{q}{m}+1\right)\kappa}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\sigma^2}{\gamma^2\mu\tau} + \frac{(q+1)\,\sigma^2}{\mu\left(\frac{q}{m}+1\right)\tau m}\right) \qquad (43)
\end{aligned}$$

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

$$1 \geq \tau^2L^2\eta^2 + \left(\frac{q}{m}+1\right)\eta\gamma L\tau \qquad (44)$$

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:*

$$\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq\ & e^{-\frac{R}{2L\left(1+\frac{q}{m}\right)\sqrt{m\tau}}}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&+ \left[\frac{\sqrt{m}\sigma^2}{8\sqrt{\tau}\gamma^2\left(1+\frac{q}{m}\right)^2} + \frac{(1+q)\,\sigma^2}{4\left(1+\frac{q}{m}\right)\sqrt{m\tau}}\right] + \frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2 \qquad (45)
\end{aligned}$$

We note that above theorem implies that to achieve a convergence error of ϵ we need to have $R = O\left(L\,(1+q)\,\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{(q+1)^2}{m\left(\frac{q}{m}+1\right)^2\epsilon^2}\right)$.

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

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

Next rearranging Eq. (46) and replacing μ with ϕ leads to the following error bound:

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

$$+\frac{\phi}{2}\left(\|\boldsymbol{w}^*\|^2-\left\|\boldsymbol{w}^{(r)}\right\|^2\right)$$
$$\leq e^{-(\eta\gamma\phi\tau)R}\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)+\frac{1}{\phi}\left[\frac{1}{2}L^2\tau\eta^2\sigma^2+(1+q)\frac{\gamma\eta L\sigma^2}{2m}\right]+\frac{\phi}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2 \tag{47}$$

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

$$f(\boldsymbol{w}^{(R)})-f^*$$
$$\leq e^{-\frac{R}{2\left(1+\frac{q}{m}\right)L\sqrt{m\tau}}}\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)+\sqrt{m\tau}\left[\frac{\sigma^2}{8\tau\gamma^2\left(1+\frac{q}{m}\right)^2}+\frac{(1+q)\,\sigma^2}{4\left(1+\frac{q}{m}\right)\tau m}\right]+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2, \tag{48}$$

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:

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta\gamma \frac{1}{m}\sum_{j=1}^{m} Q\left(\sum_{c=0}^{\tau-1} \tilde{\boldsymbol{d}}_j^{(c,r)}\right) = \boldsymbol{w}^{(r)} - \gamma \frac{1}{m}\sum_{j=1}^{m} \eta Q\left(\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\right) \tag{49}$$

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:*

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

Proof. First, note that the expression on the left hand side of (50) can be upper bounded by

$$\begin{aligned}
&\mathbb{E}_\xi\mathbb{E}_Q\left[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\|^2\right] \\
&\overset{①}{=} \eta^2\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\underbrace{Q\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)} - \boldsymbol{\Delta}_j^{(r)}\right)\right)}_{\tilde{\mathbf{g}}_Q^{(r)}}\|^2 + G_q\Big] \\
&= \eta^2\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\underbrace{Q\left(\overbrace{\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r)}\right)}^{\tilde{\mathbf{g}}^{(r)}}\right)}_{\tilde{\mathbf{g}}_Q^{(r)}}\|^2 + G_q\Big] \\
&= \eta^2\mathbb{E}_\xi\left[\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)} - \mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]\|^2\right] + \|\mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]\|^2\right] + \eta^2 G_q \\
&= \eta^2\mathbb{E}_\xi\left[\mathbb{E}_Q\left[\left\|\tilde{\mathbf{g}}_Q^{(r)} - \tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q \\
&\leq \eta^2\mathbb{E}_\xi\left[q\left\|\tilde{\mathbf{g}}^{(r)}\right\|^2 + \left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q \\
&= (q+1)\eta^2\mathbb{E}_\xi\left[\left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] + \eta^2 G_q \\
&= (q+1)\eta^2\mathbb{E}_\xi\left[\left\|\tilde{\mathbf{g}}^{(r)} - \mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2\right] + (q+1)\eta^2\left\|\mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2 + \eta^2 G_q
\end{aligned} \tag{51}$$

where ① comes from Assumption 2.

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

$$\mathbb{E}\left[\left[\|\tilde{\mathbf{g}}^{(r)} - \mathbf{g}^{(r)}\|^2\right]\right] \leq \frac{\tau\eta^2\sigma^2}{m} \tag{52}$$

To prove this claim, note that

$$\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(t)}-\mathbf{g}^{(t)}\right\|^2\right] \overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\left[\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|^2\right]$$
$$\overset{②}{=} \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\left[\left\|\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right]\right\|^2\right]$$
$$= \frac{1}{m^2}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right\|^2\right]$$
$$\overset{③}{\leq} \frac{1}{m^2}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\sigma^2 \tag{53}$$
$$= \frac{\tau\sigma^2}{m} \tag{54}$$

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 (52) into the last expression in (51) to obtain

$$\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{\Delta}_j^{(r)}\right)\|^2\Big]$$
$$\leq (q+1)\eta^2\tau\frac{\sigma^2}{m}+(q+1)\eta^2\left\|\mathbb{E}_\xi\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2+\eta^2G_q \tag{55}$$

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

$$\left\|\mathbb{E}\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|^2 = \|\mathbf{g}^{(r)}\|^2$$
$$\leq \left\|\frac{1}{m}\sum_{j=1}\left[\sum_{c=0}^{\tau-1}g_j^{(c,r)}\right]\right\|^2$$
$$\overset{①}{\leq} \left\|\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right\|^2$$
$$\overset{②}{\leq} \tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}g_j^{(c,r)}\right\|^2$$
$$= \tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 \tag{56}$$

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 (55) implies that

$$\mathbb{E}_\xi\mathbb{E}_Q\Big[\|\frac{1}{m}\sum_{j=1}^{m}\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{\Delta}_j^{(r)}\right)\|^2\Big]$$
$$\leq (q+1)\eta^2\tau\frac{\sigma^2}{m}+(q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}^{(r)}\right\|^2+\eta^2G_q, \tag{57}$$

and the proof is complete. □

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*

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

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

$$
\begin{aligned}
&-\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\mathbb{E}_Q\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}^{(r)}\right\rangle\right]\\
&=-\mathbb{E}_{\{\xi_1^{(t)},\ldots,\xi_m^{(t)}|\boldsymbol{w}_1^{(t)},\ldots,\boldsymbol{w}_m^{(t)}\}}\left[\left\langle\nabla f(\boldsymbol{w}^{(r)}),\eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right\rangle\right]\\
&=-\left\langle\nabla f(\boldsymbol{w}^{(r)}),\eta\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\tilde{\mathbf{g}}_j^{(c,r)}\right]\right\rangle\\
&=-\eta\sum_{c=0}^{\tau-1}\left\langle\nabla f(\boldsymbol{w}^{(r)}),\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\rangle\\
&\overset{①}{=}\frac{1}{2}\eta\sum_{c=0}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2-\|\frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2+\|\nabla f(\boldsymbol{w}^{(r)})-\frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2\right]\\
&\overset{②}{\leq}\frac{1}{2}\eta\sum_{c=0}^{\tau-1}\frac{1}{m}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2-\|\frac{1}{m}\sum_{j=1}^{m}\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2+L^2\frac{1}{m}\sum_{j=1}^{m}\|\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(c,r)}\|_2^2\right],
\end{aligned}
\tag{59}
$$

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.

□

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

$$
\begin{aligned}
&\frac{1}{R}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right\|^2\\
&\quad\leq 36\eta^2\tau^2\sigma^2+\frac{8\eta^2}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&\qquad+10\eta^2(\eta\gamma)^2(q+1)L^2\left[\frac{\tau^4}{R}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r-1)}\right\|^2\right]+\tau^4\frac{\sigma^2}{m}+\tau^3G_q\right]\\
&\qquad+\frac{20\eta^2\tau^2}{R}\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2.
\end{aligned}
\tag{60}
$$

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`(τ, η, γ), *for all* $0 \leq t \leq R\tau - 1$, *under Assumptions 1, 2, 4 and 5 and if the learning rate satisfies*

$$1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - L\eta\gamma\tau(q+1) \geq 0 \quad \& \quad 30\eta^2L^2\tau^2 \leq 1 \tag{61}$$

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

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 \leq{}& \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right)}{\tau\eta\gamma R} + \frac{(q+1)\gamma L\eta\sigma^2}{m} + 36\eta^2L^2\tau\sigma^2 + 10\eta^2L^4\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} \\ &+ 10\eta^2L^4\tau^2(\eta\gamma)^2(q+1)G_q + \frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)}) - f_j(\boldsymbol{w}_j^{(*)})\right] + \frac{16\eta^3L^2\tau^2}{R}\sigma^2 \\ &+ \frac{32\eta^2L^3\tau^2}{R}\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) + \frac{\gamma\eta L}{\tau}G_q\end{aligned} \tag{62}$$

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:

$$\begin{aligned}\delta_j^{(r)} &= \delta_j^{(r-1)} + \frac{1}{\tau}\left(\boldsymbol{\Delta}_q^{(k)} - \boldsymbol{\Delta}_{j,q}^{(k)}\right) \\ &= \delta_j^{(r-1)} + \frac{1}{\tau}\left(\frac{1}{m}\sum_{j=1}^{m}Q\left(-\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \delta_j^{(r-1)}\right)\right) + Q\left(\sum_{c=0}^{\tau-1}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \delta_j^{(r-1)}\right)\right)\right)\end{aligned} \tag{63}$$

Therefore, we have

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

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

$$\mathbb{E}_Q\left[\tilde{\boldsymbol{d}}_{jq}^{(c,r)}\right] = \mathbb{E}_Q\left[\tilde{\mathbf{g}}_j^{(c,r)} - \delta_j^{(r)}\right] = \tilde{\mathbf{g}}_j^{(c,r)} + \frac{1}{\tau}\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)} - \sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right) = \tilde{\boldsymbol{d}}_j^{(c,r)} \tag{65}$$

We would like to also highlight that

$$-\eta Q\left(\frac{\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(\tau,r)}}{\eta}\right)=-\eta Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jQ}^{(c,r)}\right) \tag{66}$$

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:

$$\mathbb{E}_Q\left[\tilde{\mathbf{g}}_Q^{(r)}\right]=\frac{1}{m}\sum_{j=1}\left[-\eta\mathbb{E}_Q\left[Q\left(\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_{jq}^{(c,r)}\right)\right]\right]=\frac{1}{m}\sum_{j=1}\left[-\eta\left(\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right)\right]\triangleq\tilde{\mathbf{g}}^{(r)} \tag{67}$$

Then following the L-smoothness gradient assumption on global objective, by using $\tilde{\mathbf{g}}^{(r)}$ in inequality (66) we have:

$$f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\leq-\gamma\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}_Q^{(r)}\rangle+\frac{\gamma^2L}{2}\|\tilde{\mathbf{g}}_Q^{(r)}\|^2 \tag{68}$$

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

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_Q\left[f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\right]\right]&\leq-\gamma\mathbb{E}\left[\mathbb{E}_Q\left[\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}_Q^{(r)}\rangle\right]\right]+\frac{\gamma^2L}{2}\mathbb{E}\left[\mathbb{E}_Q\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right]\\&\overset{①}{=}-\gamma\underbrace{\mathbb{E}\left[\left[\langle\nabla f(\boldsymbol{w}^{(r)}),\tilde{\mathbf{g}}^{(r)}\rangle\right]\right]}_{\text{(I)}}+\frac{\gamma^2L}{2}\underbrace{\mathbb{E}\left[\mathbb{E}_Q\left[\|\tilde{\mathbf{g}}_Q^{(r)}\|^2\right]\right]}_{\text{(II)}}\end{aligned} \tag{69}$$

where ① follows from Eq. (67). Next, by plugging back the results in Lemma E.1, Lemma E.2, and Lemma E.3 we obtain

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

which leads to

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

$$
\begin{aligned}
&\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{71}
$$

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

$$
1 - 10\eta^2(\eta\gamma)^2(q+1)L^4\tau^4 - (q+1)L\eta\gamma\tau \geq 0. \tag{72}
$$

Rearranging Eq. (71) we obtain:

$$
\begin{aligned}
&\left(1 - 20\eta^2L^2\tau^2\right)\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 \\
&\quad \leq \frac{2\left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^*)\right)}{\tau\eta\gamma R} + \frac{(q+1)\gamma L\eta\sigma^2}{m} + 36\eta^2L^2\tau\sigma^2 + 10\eta^2L^4\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} \\
&\quad + 10\eta^2L^4\tau^2(\eta\gamma)^2(q+1)G_q + \underbrace{\frac{8\eta^2L^2}{m\tau R}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2}_{\text{(IV)}} + \frac{\gamma\eta L}{\tau}G_q,
\end{aligned}
\tag{73}
$$

and the claim follows.

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

$$
\frac{8\eta^2L^2}{m\tau R}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2
$$

$$
\begin{aligned}
&= \frac{8\eta^2L^2\tau}{m\tau R}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&\overset{①}{\leq} \frac{8\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right\|^2\\
&\leq \frac{16\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[\left\|\mathbf{g}_j^{(c,0)}\right\|^2+\left\|\mathbf{g}^{(0)}\right\|^2\right]\\
&= \frac{16\eta^2L^2\tau^2}{mR}\sum_{j=1}^{m}\left[\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)}\right\|^2\right]+\frac{16\eta^2L^2\tau}{mR}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{②}{\leq} \frac{16\eta^2L^2\tau^2}{mR}\sum_{j=1}^{m}\left[\frac{2\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]}{\eta\tau}+\eta\sigma^2\right]+\frac{16\eta^2L^2\tau^2}{R}\left\|\mathbf{g}^{(0)}\right\|^2\\
&= \frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]+\frac{16\eta^3L^2\tau^2}{R}\sigma^2+\frac{16\eta^2L^2\tau^2}{R}\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{③}{\leq} \frac{32\eta L^2\tau}{mR}\sum_{j=1}^{m}\left[f_j(\boldsymbol{w}_j^{(0)})-f_j(\boldsymbol{w}_j^{(*)})\right]+\frac{16\eta^3L^2\tau^2}{R}\sigma^2+\frac{32\eta^2L^3\tau^2}{R}\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)
\end{aligned}
\tag{74}
$$

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 [8], 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 [13, 57] for more details). This completes the proof.

□

Remark 8. *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. (61) is satisfied simultaneously, we need to have*

$$
1 \geq \frac{10m^2\tau^2}{\gamma^2R^2(q+1)}+\sqrt{\frac{m\tau\,(q+1)}{R}} \tag{75}
$$

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:

$$
R \geq m\tau\left(\frac{q+1}{2}\right) \tag{76}
$$

We note that if we solve this inequality such as Eq. (33) 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. (62) 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 [26] and [47] with tuned global and local learning rates, showing that we can archive the error ϵ 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 [26] and [36], which shows that obtained rate is tight. We highlight that the communication complexity of our algorithm is better than [26] in terms of number of bits per iteration as we do not use additional control variable.

Remark 9. *We note that the conditions in Eq. (61) can be rewritten as*

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

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`(τ, η, γ), for all $0 \leq t \leq R\tau - 1$, under Assumptions 1, 2, 4, 5 and 6 and if the learning rate satisfies*

$$1 - (q+1)L\eta\gamma\tau - \frac{10(q+1)\eta^2\tau^4L^4(\eta\gamma)^2}{1 - \mu\tau\gamma\eta + 20\mu\gamma\eta^3L^2\tau^3} \geq 0 \quad \& \quad 30\eta^2L^2\tau^2 \leq 1 \tag{78}$$

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

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq \left(1 - \frac{\mu\eta\gamma\tau}{3}\right)^R \Big(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big) \\
&\quad + \frac{3}{\mu}\Big[L^2 18\eta^2\tau\sigma^2 + \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\Big\|\Big(\boldsymbol{w}_j^{(0,0)} - \boldsymbol{w}_j^{(*)}\Big)\Big\|_2^2 + 16L^3\tau^2\eta^2\Big(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big) \\
&\quad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big].
\end{aligned} \tag{79}$$

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:*

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2 \\
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\eta\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2 \\
&\leq 36\eta^2\tau^2\sigma^2 + \frac{8\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\right\|^2 \\
&\quad + 10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r-1)}\right\|^2 \\
&\quad + 10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + 20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2
\end{aligned} \tag{80}$$

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

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(r)})\Big]$$

$$
\begin{aligned}
&\leq \frac{1}{2}\gamma\eta \sum_{c=0,r}^{\tau-1}\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \frac{L^2}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2\right] \\
&\quad + \frac{\gamma^2 L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2 G_q\right] \\
&= -\frac{\tau\gamma\eta}{2}\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 + \frac{1}{2}\gamma\eta\sum_{c=0,r}^{\tau-1}\left[-\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \frac{L^2}{m}\sum_{j=1}^{m}\mathbb{E}\left\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\right\|^2\right] \\
&\quad + \frac{\gamma^2 L}{2}\left[(q+1)\eta^2\tau\frac{\sigma^2}{m} + (q+1)\eta^2\tau\sum_{c=0}^{\tau-1}\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_j^{(c,r)}\right\|^2 + \eta^2 G_q\right] \qquad (81)
\end{aligned}
$$

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

$$+ L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Big\|^2$$
$$+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{\eta^2\gamma^2L}{2}\frac{(q+1)\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2} \tag{82}$$

where ① follows from

$$1 - (q+1)L\eta\gamma\tau \geq 0 \tag{83}$$

Next Eq. (82) leads us to

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)}) - f(\boldsymbol{w}^{(*)})\Big] = a_{r+1}$$
$$\leq \big(1 - \mu\tau\gamma\eta + 20\mu\gamma\eta^3L^2\tau^3\big)\Big(f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(*)})\Big) + \frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\|\sum_{j=1}^{m}\frac{1}{m}\nabla f_j(\boldsymbol{w}_j^{(c,r-1)})\|_2^2$$
$$+ L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Big\|^2$$
$$+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2}$$
$$\overset{①}{=} \Delta a_r + \frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-1} + c$$
$$\overset{(a)}{\leq} \Delta\Big(\Delta a_{r-1} + \frac{\eta\gamma}{2}\left(1 - (q+1)L\eta\gamma\tau\right)e_{r-1} + \frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-2} + c\Big) + \frac{1}{2}\gamma^2\eta^2 10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-1} + c$$
$$= \Delta^2 a_{r-1} + \frac{\eta\gamma}{2}\left(\Delta - \Delta(q+1)L\eta\gamma\tau - 10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)\right)e_{r-1} + \frac{\Delta\gamma^2\eta^2}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-2}$$
$$+ (\Delta+1)c$$
$$\overset{(b)}{\leq} \Delta^2 a_{r-1} + \frac{\Delta\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-2} + c\Delta + c$$
$$= \Delta\Big(\Delta a_{r-1} + \frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-2} + c\Big) + c$$
$$\overset{(d)}{\leq} \Delta\Big(\Delta\Big(\Delta a_{r-2} + \frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{r-3} + c\Big) + c\Big) + c$$
$$\overset{(e)}{\leq} \Delta^r a_0 + \Delta^{r-1}\frac{\gamma\eta}{2}10\eta^2L^4\tau^4(\eta\gamma)^2(q+1)e_{-1} + \left(\Delta^{r-1} + \Delta^{r-2} + \ldots + 1\right)c$$
$$\overset{(f)}{=} \Delta^r a_0 + \left(\frac{1-\Delta^r}{1-\Delta}\right)c$$
$$\overset{①}{=} \Delta^r\Big(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big)$$
$$+ \frac{1-\Delta^r}{1-\Delta}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Big\|^2$$
$$+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2}\Big]$$
$$\leq \Delta^r\Big(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\Big)$$
$$+ \frac{1}{1-\Delta}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Big\|^2$$
$$+ 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2}\Big] \tag{84}$$

where ① holds because of $\Delta = 1 - \mu\tau\gamma\eta + 20\mu\gamma\eta^3L^2\tau^3$, and the following short hand notations:

$$a_r = \mathbb{E}\Big[f(\boldsymbol{w}^{(r)}) - f(\boldsymbol{w}^{(*)})\Big]$$

$$
\begin{aligned}
e_r &= \sum_{c=0,r}^{\tau-1} \| \sum_{j=1}^{m} \frac{1}{m} \nabla f_j(\boldsymbol{w}_j^{(c,r)}) \|_2^2 \\
c &= L^2 18 \gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta \frac{4\eta^2}{m} \sum_{j=1}^{m}\sum_{c=0}^{\tau-1} \left\| \sum_{c=0,r=0}^{\tau-1} \left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)} \right) \right\|^2 \\
&\quad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{\eta^2\gamma^2 L}{2}\frac{(q+1)\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2 G_q}{2}
\end{aligned} \tag{85}
$$

(a) comes from reapplying the recursion. (b) is due to the condition

$$
1 - (q+1)L\eta\gamma\tau - \frac{10(q+1)\eta^2\tau^4L^4(\eta\gamma)^2}{\Delta} = 1 - (q+1)L\eta\gamma\tau - \frac{10(q+1)\eta^2\tau^4L^4(\eta\gamma)^2}{1-\mu\tau\gamma\eta+20\mu\gamma\eta^3L^2\tau^3} \geq 0 \tag{86}
$$

(d) comes from one step reapplying of recursion. (e) holds by repeating the recursion under the same condition of learning rate for $r-1$ times. Finally, (f) follows from $e_{-1}=0$, which leads the following bound:

$$
\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \\
&\leq \big(1 - \mu\eta\gamma\tau\left(1 - 20\eta^2L^2\tau^2\right)\big)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{1}{\mu\eta\gamma\tau\left(1-20\eta^2L^2\tau^2\right)}\Big[L^2 18\gamma\eta\eta^2\tau^2\sigma^2 + L^2\gamma\eta\frac{4\eta^2}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad + 5L^4\gamma\eta\eta^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\gamma\eta\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta^2\gamma^2L}{2}\frac{\tau\sigma^2}{m} + \frac{L\eta^2\gamma^2G_q}{2}\Big] \\
&= \big(1 - \mu\eta\gamma\tau\left(1 - 20\eta^2L^2\tau^2\right)\big)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{1}{\mu\left(1-20\eta^2L^2\tau^2\right)}\Big[L^2 18\eta^2\tau\sigma^2 + \frac{4L^2\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 \\
&\qquad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big] \\
&\overset{①}{\leq} \left(1 - \frac{\mu\eta\gamma\tau}{3}\right)^r \left(f(\boldsymbol{w}^{(0)}) - f(\boldsymbol{w}^{(*)})\right) \\
&\quad + \frac{3}{\mu}\Big[L^2 18\eta^2\tau\sigma^2 + \underbrace{L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2}_{\text{(V)}} \\
&\qquad + 5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m} + 5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q + \frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m} + \frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned} \tag{87}
$$

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

Finally we continue with bounding term (V):

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

$$
\begin{aligned}
&= \frac{8L^2\eta^2}{m}\sum_{j=1}^{m}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(*)}\right)\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)}-\mathbf{g}^{(0)}\right\|^2\\
&\le \frac{8L^2\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)}-\mathbf{g}^{(0)}\right\|^2\\
&\overset{①}{\le} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}_j^{(*)}-\mathbf{g}^{(0)}\right\|^2\\
&\overset{②}{=} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}_j^{(*)}\right\|^2 + \frac{8L^2\tau^2\eta^2}{m}\sum_{j=1}^{m}\left\|\mathbf{g}^{(0)}\right\|^2\\
&= \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[\left\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}_j^{(*)}\right\|^2\right] + 8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{③}{\le} \frac{8L^4\eta^2\tau}{m}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\left[(1-2\mu\eta\,(1-\eta L))^c\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + \frac{\eta\sigma^2}{\mu(1-\eta L)}\right] + 8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{④}{\le} \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\left[\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + \frac{\eta\sigma^2}{\mu(1-\eta L)}\right] + 8L^2\tau^2\eta^2\left\|\mathbf{g}^{(0)}\right\|^2\\
&\overset{⑤}{\le} \frac{8L^4\eta^2\tau^2}{m}\sum_{j=1}^{m}\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + \frac{8L^4\eta^3\tau^2\sigma^2}{\mu(1-\eta L)} + 16L^3\tau^2\eta^2\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)
\end{aligned}
\tag{88}
$$

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 $\left\|\boldsymbol{w}_j^{(t,0)}-\boldsymbol{w}_j^{(*)}\right\|^2 \le (1-2\mu\eta\,(1-\eta L))^t\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2 + \frac{\eta\sigma^2}{\mu(1-\eta L)}$ [44], ④ holds due to the choice of learning rate η such that $(1-2\mu\eta\,(1-\eta L))^c \le 1$, and finally ⑤ is due to smoothness assumption which implies $\left\|\mathbf{g}^{(0)}\right\|^2 \le 2L\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)$ 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 $\gamma\ge\sqrt{m\tau}$ in Eq. (79) which incurs:*

$$
\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big] \le{}& \left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)e^{-\left(\frac{R}{6(q+1)\kappa}\right)}\\
&+\frac{3}{\mu}\Big[\frac{4.5\sigma^2}{\tau\gamma^2(q+1)^2}+\frac{2L^2}{(q+1)^2\gamma^2 m}\sum_{j=1}^{m}\left\|\left(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\right)\right\|_2^2+\frac{2L\left(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right)}{(q+1)^2\gamma^2}\\
&+\frac{\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau-0.5\right)}+\frac{5}{16(q+1)^3\tau\gamma^2}\frac{\sigma^2}{m}+\frac{5}{16(q+1)^3\tau^2\gamma^2}G_q\\
&+\frac{1}{4\tau}\frac{\sigma^2}{m}+\frac{G_q}{4(q+1)\tau^2}\Big]
\end{aligned}
\tag{89}
$$

From Eq. (89) we can see that to attain an ϵ-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`(τ,η,γ) *(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*

$$1-(q+1)L\eta\gamma\tau-\frac{10(q+1)\eta^2\tau^4L^4(\eta\gamma)^2}{1-\mu\tau\gamma\eta+20\mu\gamma\eta^3L^2\tau^3}\geq 0\quad\&\quad 30\eta^2L^2\tau^2\leq 1 \tag{90}$$

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:

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big]\\
&\leq e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)\\
&\quad+\Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}}+\frac{6\sqrt{m\tau}L^2}{m(q+1)^2\gamma^2}\sum_{j=1}^{m}\Big\|\Big(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\Big)\Big\|_2^2+\frac{12L\sqrt{m\tau}}{\gamma^2(q+1)^2}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)\,.\\
&\quad+\frac{3\sqrt{m\tau}\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau-0.5\right)}+\frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}}+\frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2}+\frac{3\sigma^2}{4\sqrt{m\tau}}+\frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big]\\
&\quad+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned}\tag{91}$$

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

$$\begin{aligned}
\tilde{f}(\boldsymbol{w}^{(R)},\phi)-\tilde{f}(\boldsymbol{w}^{(*)},\phi)&=f(\boldsymbol{w}^{(r)})+\frac{\lambda}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2-\left(f(\boldsymbol{w}^{(*)})+\frac{\lambda}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2\right)\\
&\leq\left(1-\frac{\eta\gamma\phi\tau}{3}\right)^R\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)\\
&\quad+\frac{3}{\phi}\Big[L^2 18\eta^2\tau\sigma^2+L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\Big\|^2\\
&\quad+5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}+5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q+\frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m}+\frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned}\tag{92}$$

Next rearranging Eq. (92) and replacing μ with ϕ, and using the short hand notation of

$$\begin{aligned}
\mathcal{A}(\eta)\triangleq\Big[&L^2 18\eta^2\tau\sigma^2+L^2\frac{4\eta^2}{m\tau}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\Big\|^2\\
&+5L^4\eta^2\tau^3(\eta\gamma)^2(q+1)\frac{\sigma^2}{m}+5L^2\eta^2L^2\tau^2(\eta\gamma)^2(q+1)G_q+\frac{(q+1)\eta\gamma L}{2}\frac{\sigma^2}{m}+\frac{L\eta\gamma G_q}{2\tau}\Big]
\end{aligned}\tag{93}$$

leads to the following error bound:

$$\begin{aligned}
\tilde{f}(\boldsymbol{w}^{(R)},\phi)-f^*&\leq\left(1-\frac{\eta\gamma\phi\tau}{3}\right)^R\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)+\frac{3}{\phi}\mathcal{A}(\eta)+\frac{\phi}{2}\left(\left\|\boldsymbol{w}^{(*)}\right\|^2-\left\|\boldsymbol{w}^{(r)}\right\|^2\right)\\
&\leq e^{-\left(\frac{\eta\gamma\phi\tau}{3}\right)R}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)+\frac{3}{\phi}\mathcal{A}(\eta)+\frac{\phi}{2}\left\|\boldsymbol{w}^{(*)}\right\|^2
\end{aligned}\tag{94}$$

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:

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

$$+\Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}}+\frac{3}{\sqrt{m}\,(q+1)^2\,\gamma^2\tau^{2.5}}\sum_{j=1}^{m}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\Big\|^2$$
$$+\frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}}+\frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2}+\frac{3\sigma^2}{4\sqrt{m\tau}}+\frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big]$$
$$+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2 \tag{95}$$

Finally, using Eq. (88) we obtain the bound:

$$\tilde{f}(\boldsymbol{w}^{(R)},\phi)-f^*\leq e^{-\frac{R}{6(1+q)L\sqrt{m\tau}}}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)$$
$$+\Big[\frac{13.5\sqrt{m}\sigma^2}{(q+1)^2\gamma^2\sqrt{\tau}}+\frac{6\sqrt{m\tau}L^2}{(q+1)^2\gamma^2 m}\sum_{j=1}^{m}\Big\|\Big(\boldsymbol{w}_j^{(0,0)}-\boldsymbol{w}_j^{(*)}\Big)\Big\|_2^2+\frac{12L\sqrt{m\tau}}{\gamma^2(q+1)^2}\Big(f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big)$$
$$+\frac{3\sqrt{m\tau}\kappa\sigma^2}{(q+1)^2\gamma^2\left((q+1)\gamma\tau-0.5\right)}+\frac{15\sigma^2}{16(q+1)^3\gamma^2\sqrt{m\tau}}+\frac{15G_q\sqrt{m}}{16(q+1)^3\tau^{1.5}\gamma^2}+\frac{3\sigma^2}{4\sqrt{m\tau}}+\frac{3\sqrt{m}G_q}{4(q+1)\tau^{1.5}}\Big]$$
$$+\frac{1}{2\sqrt{m\tau}}\left\|\boldsymbol{w}^{(*)}\right\|^2 \tag{96}$$

□

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 ϵ 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:*

$$\begin{aligned}\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\left(\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(c,r)}\right)\right\|^2 &= \frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\mathbb{E}\left\|\sum_{c=0,r}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\right\|^2 \\ &\leq 36R\eta^2\tau^2\sigma^2 + 8\eta^2 C + 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \\ &\quad + \frac{10\eta^2L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2 \end{aligned} \tag{97}$$

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 $r \geq 1$:*

$$\begin{aligned}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\tilde{d}_j^{(c,r)}\right\|^2 &\leq 18\sigma^2\tau + \frac{1}{p}\sum_{j=1}^{p}\Big[6L^2\tau\Big[\sum_{c=0,r}^{\tau-1}\left\|\left[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\right]\right\|^2 + \frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2 \\ &\quad + \sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}_j^{(c,r-1)}\right\|^2 + \sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\left\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2 \\ &\quad + \sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r-1)}-\boldsymbol{w}^{(r-1)}\right\|^2 + \frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\Big] \end{aligned} \tag{98}$$

Proof.

$$\begin{aligned}&\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\tilde{d}_j^{(c,r)}\right\|^2 \\ &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left\|\sum_{c=0}^{\tau-1}\left[\tilde{\mathbf{g}}_j^{(c,r)} + \frac{1}{\tau}\left(\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)} - \sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\right]\right\|^2 \\ &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)} + \mathbf{g}_j^{(c,r)}\right) \\ &\quad + \frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \mathbf{g}_j^{(c,r-1)} + \mathbf{g}_j^{(c,r-1)}\right) - \mathbf{g}_j^{(c,r-1)} + \mathbf{g}_j^{(c,r-1)} - \tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2 \\ &\leq \frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\underbrace{\Big\|\sum_{c=0}^{\tau-1}\Big[\left(\tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)}\right) + \frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\left(\tilde{\mathbf{g}}_j^{(c,r-1)} - \mathbf{g}_j^{(c,r-1)}\right) + \mathbf{g}_j^{(c,r-1)} - \tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2}_{\text{(I)}} \\ &\quad + \frac{2}{p}\sum_{j=1}^{p}\underbrace{\Big\|\sum_{c=0}^{\tau-1}\Big[\mathbf{g}_j^{(c,r)} + \frac{2}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(c,r-1)}\right)\Big]\Big\|^2}_{\text{(II)}} \end{aligned} \tag{99}$$

□

We first bound the term (I) in Eq. (99) with the following lemma:

Lemma F.3.

$$\frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\Big(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\Big)+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\Big(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\Big)+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2\leq 18\sigma^2\tau \quad (100)$$

Proof.

$$\begin{aligned}
&\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big[\Big(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\Big)+\frac{1}{\tau}\sum_{c=0}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\Big(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\Big)+\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\right)\Big]\Big\|^2\\
&\leq \mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\Big)\Big\|^2+\Big\|\frac{1}{\tau}\sum_{c=0}^{\tau-1}\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\Big)\Big\|^2+\Big\|\frac{1}{\tau}\sum_{c=0}^{\tau-1}\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\Big)\Big\|^2\\
&=3\left[\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\Big)\Big\|^2\right]\\
&\overset{①}{=}3\left[\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\Big(\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\Big)\Big\|^2+\sum_{c=0}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\Big(\tilde{\mathbf{g}}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\Big)\Big\|^2+\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\Big(\mathbf{g}_j^{(c,r-1)}-\tilde{\mathbf{g}}_j^{(c,r-1)}\Big)\Big\|^2\right]\\
&\leq\tau\left(\sigma^2+\sigma^2+\sigma^2\right)\\
&=9\sigma^2\tau
\end{aligned} \quad (101)$$

where ① follows from Assumption 4. □

We bound the term (II) in Eq. (99) as follows:

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

$$\Big\|\sum_{c=0,r}^{\tau-1}\Big[\mathbf{g}_j^{(c,r)}+\frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\left(\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(c,r-1)}\right)\Big]\Big\|^2 \quad (102)$$

$$\begin{aligned}
&\leq 5L^2\Big[\tau\sum_{c=0,r}^{\tau-1}\Big\|\Big[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\Big]\Big\|^2+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\Big\|^2\\
&\quad+\tau\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\tau\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2\\
&\quad+\tau\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\tau\frac{1}{L^2}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^2\Big]
\end{aligned} \quad (103)$$

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

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

$$
\begin{aligned}
&= \Big\| \sum_{c=0}^{\tau-1} \Big[\mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} + \mathbf{g}_j^{(r)} - \frac{1}{\tau} \sum_{c=0}^{\tau-1} \mathbf{g}_j^{(r-1)} \\
&\quad + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \left(\mathbf{g}_j^{(r-1)} - \mathbf{g}_j^{(c,r-1)} \right) + \frac{1}{\tau} \sum_{c=0}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) + \mathbf{g}^{(r-1)} \Big] \Big\|^2 \\
&\leq 5 \Big[\Big\| \sum_{c=0,r}^{\tau-1} \left[\mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \right] \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \left(\mathbf{g}_j^{(r)} - \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \left(\mathbf{g}_j^{(r-1)} - \mathbf{g}_j^{(c,r-1)} \right) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big] \\
&\overset{①}{=} 5 \Big[\Big\| \sum_{c=0,r}^{\tau-1} \left[\mathbf{g}_j^{(c,r)} - \mathbf{g}_j^{(r)} \right] \Big\|^2 + \Big\| \sum_{c=0,r}^{\tau-1} \frac{1}{\tau} \sum_{c=0,r-1}^{\tau-1} \left(\mathbf{g}_j^{(r)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \left(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 + \Big\| \frac{1}{\tau} \sum_{c=0,r}^{\tau-1} \sum_{c=0,r-1}^{\tau-1} \frac{1}{p} \sum_{j=1}^{p} \left(\mathbf{g}_j^{(c,r-1)} - \mathbf{g}_j^{(r-1)} \right) \Big\|^2 \\
&\quad + \Big\| \sum_{c=0,r-1}^{\tau-1} \mathbf{g}^{(r-1)} \Big\|^2 \Big]
\end{aligned} \tag{104}
$$

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. (104):

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

$$\leq 5L^2\Big[\tau\sum_{c=0,r}^{\tau-1}\Big\|\Big[\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\Big]\Big\|^2+\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\Big\|^2$$
$$+\tau\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\tau\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big\|\boldsymbol{w}_j^{(c,r-1)}-\boldsymbol{w}^{(r-1)}\Big\|^2+\frac{\tau}{L^2}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^2\Big] \tag{105}$$

where ② is due to the fact that

$$\Big\|\frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big(\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(r-1)}\Big)\Big\|^2+\Big\|\frac{1}{\tau}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big(\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(r-1)}\Big)\Big\|^2$$
$$=\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\Big(\mathbf{g}_j^{(r)}-\mathbf{g}_j^{(r-1)}\Big)\Big\|^2+\Big\|\sum_{c=0,r-1}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\Big(\mathbf{g}_j^{(c,r-1)}-\mathbf{g}_j^{(r-1)}\Big)\Big\|^2, \tag{106}$$

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:*

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r)}\Big\|^2\leq\frac{4}{p}\sum_{j=1}^{p}\Bigg[\tau\sigma^2+\tau L^2\sum_{c=0}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(c,0)}\Big)\Big\|^2+\tau\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Bigg] \tag{107}$$

Proof. For $r=0$ we have:

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,0)}\Big\|^2$$
$$=\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,0)}\Big\|^2$$
$$=\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}+\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}+\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}+\mathbf{g}^{(0)}\Big)\Big\|^2$$
$$\leq\frac{4}{p}\sum_{j=1}^{p}\Bigg[\mathbb{E}\Big\|\sum_{c=0}^{\tau-1}\Big(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\mathbf{g}^{(0)}\Big\|^2\Bigg]$$
$$\overset{①}{=}\frac{4}{p}\sum_{j=1}^{p}\Bigg[\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\Big(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\mathbf{g}^{(0)}\Big\|^2\Bigg]$$
$$\leq\frac{4}{p}\sum_{j=1}^{p}\Bigg[\sum_{c=0}^{\tau-1}\mathbb{E}\Big\|\Big(\tilde{\mathbf{g}}_j^{(c,0)}-\mathbf{g}_j^{(c,0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\Big)\Big\|^2+\tau\sum_{c=0}^{\tau-1}\Big\|\Big(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\Big)\Big\|^2+\tau\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Bigg]$$
$$\leq\frac{4}{p}\sum_{j=1}^{p}\Bigg[\sum_{c=0}^{\tau-1}\sigma^2+\tau\sum_{c=0}^{\tau-1}\Big\|\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}_j^{(0)}\Big)\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(0)}-\mathbf{g}^{(0)}\Big)\Big\|^2+\tau\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Bigg]$$
$$=\frac{4}{p}\sum_{j=1}^{p}\Bigg[\tau\sigma^2+\tau L^2\sum_{c=0}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,0)}-\boldsymbol{w}^{(c,0)}\Big\|^2+\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(c,0)}\Big)\Big\|^2+\tau\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Bigg] \tag{108}$$

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. (99) as shown below.

Proof.

$$\boldsymbol{w}_j^{(c,r)} = \boldsymbol{w}_j^{(c-1,r)} - \eta \tilde{\boldsymbol{d}}_j^{(c,r)} = \ldots = \boldsymbol{w}^{(r)} - \eta \sum_{k=0}^{c-1} \tilde{\boldsymbol{d}}_j^{(c,r)} \tag{109}$$

Now we can write:

$$\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\\
&=\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]\\
&\le\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]\\
&\quad+\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\\
&\quad+\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\\
&\quad+\frac{\tau}{L^2}\sum_{c=0,r-1}^{\tau-1}\left\|\mathbf{g}^{(r-1)}\right\|^2\Big]\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\\
&\quad+\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]\\
&\quad+\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\\
&\quad+\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\\
&\quad+\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]\\
&\quad+\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\\
&\quad+\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\\
&\quad+\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)
\end{aligned}\tag{110}$$

Now we continue with bounding Eq. (110) with further simplification as follows:

$$
\begin{aligned}
&= \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}\Big\|\boldsymbol{w}_j^{(c,0)} - \boldsymbol{w}^{(c,0)}\Big\|^2 + \frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big)\Big\|^2 + 4\tau^2\sum_{c=0,r=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Big] \\
&\quad + 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}\Big\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big\|^2 + 5L^2\tau\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\Big\|^2 \\
&\quad + \frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}^{(r-1)} - \boldsymbol{w}_j^{(c,r-1)}\Big\|^2 + 5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^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}\Big\|\boldsymbol{w}_j^{(c,0)} - \boldsymbol{w}^{(c,0)}\Big\|^2 + \frac{4}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big\|^2 + 4\tau^2\sum_{c=0,r=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Big] \\
&\quad + \frac{5L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r}^{\tau-1}\Big\|\Big[\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big]\Big\|^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}\Big\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\Big\|^2 \\
&\quad + \frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r-1)} - \boldsymbol{w}^{(r-1)}\Big\|^2 + 5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^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}\Big\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big\|^2 + 5\tau^2\sum_{c=0,r=0}^{\tau-1}\Big\|\mathbf{g}^{(0)}\Big\|^2\Big] \\
&\quad + \frac{5L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\Big\|\Big[\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big]\Big\|^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}\Big\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\Big\|^2 \\
&\quad + \frac{10L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big\|^2 + 5\tau^2\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}^{(r-1)}\Big\|^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}\Big\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\Big\|^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}\Big\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\Big\|^2 \\
&\quad + \frac{15\eta^2L^2\tau^2}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\Big\|\boldsymbol{w}_j^{(c,r)} - \boldsymbol{w}^{(r)}\Big\|^2 + 10\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(r)}\Big\|^2 \qquad (111)
\end{aligned}
$$

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

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

Rearranging Eq. (111) we obtain:

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

$$\overset{①}{\leq} 36R\eta^2\tau^2\sigma^2 + \frac{36\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)} - \mathbf{g}^{(0)}\right)\right\|^2 + \frac{10\eta^2L^2\tau}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2 + 20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2 \tag{112}$$

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:*

$$\frac{1}{p}\sum_{j=1}^{p}\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\mathbb{E}_\xi\mathbb{E}_Q\left\|\boldsymbol{w}^{(r)} - \boldsymbol{w}^{(r-1)}\right\|^2 \leq \tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\right\|^2\right] + \tau^3R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p} + \tau^2(\eta\gamma)^2(q+1)RG_q \tag{113}$$

Proof.

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

$$= \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathbb{E}_\xi\Big[\Big\|\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\boldsymbol{d}}_j^{(c,r-1)}\Big]\Big\|^2+G_q\Big]$$
$$\overset{④}{=} \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathbb{E}_\xi\Big[\Big\|\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big\|^2+G_q\Big] \tag{114}$$

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. (114) as follows:

$$= \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathbb{E}_\xi\Big[\Big\|\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big\|^2\Big]+\tau^2(\eta\gamma)^2(q+1)RG_q$$
$$= \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathrm{Var}_\xi\Big(\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big)+\tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\Big\|\mathbb{E}_\xi\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Big\|^2$$
$$+\tau^2(\eta\gamma)^2(q+1)RG_q$$
$$= \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\mathrm{Var}_\xi\Big(\Big[\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big)+\tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\Big\|\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathbf{g}_j^{(c,r-1)}\Big\|^2$$
$$+\tau^2(\eta\gamma)^2(q+1)RG_q$$
$$\overset{①}{=} \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathrm{Var}_\xi\Big(\Big[\tilde{\mathbf{g}}_j^{(c,r-1)}\Big]\Big)+\tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\Big\|\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\mathbf{g}_j^{(c,r-1)}\Big\|^2$$
$$+\tau^2(\eta\gamma)^2(q+1)RG_q$$
$$\overset{②}{\le} \tau^2(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\sigma^2+\tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0,r-1}^{\tau-1}\Big\|\mathbf{g}_j^{(c,r-1)}\Big\|^2+\tau^2(\eta\gamma)^2(q+1)RG_q$$
$$= \tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}_j^{(c,r-1)}\Big\|^2+\tau^3(\eta\gamma)^2(q+1)R\frac{1}{p}\sigma^2+\tau^2(\eta\gamma)^2(q+1)RG_q \tag{115}$$

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

Finally, by plugging Lemma F.6 into Eq. (112), we obtain the following bound:

$$\sum_{r=0}^{R-1}\sum_{c=0,r}^{\tau-1}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\Big\|\boldsymbol{w}_j^{(c,r)}-\boldsymbol{w}^{(r)}\Big\|^2$$
$$\le 36R\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big\|^2$$
$$+10\eta^2L^2\tau\Big[\tau^3(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big[\Big\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\Big\|^2\Big]+\tau^3R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p}+\tau^2(\eta\gamma)^2(q+1)RG_q\Big]$$
$$+20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(r)}\Big\|^2$$
$$= 36R\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\Big\|\sum_{c=0,r=0}^{\tau-1}\Big(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\Big)\Big\|^2$$

$$+\Big[10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{r=1}^{R-1}\sum_{c=0,r-1}^{\tau-1}\Big[\Big\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\Big\|^2\Big]$$
$$+10\eta^2L^2\tau^4R(\eta\gamma)^2(q+1)\frac{\sigma^2}{p}+10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)RG_q\Big]$$
$$+20\eta^2\tau^2\sum_{r=0}^{R-1}\sum_{c=0}^{\tau-1}\Big\|\mathbf{g}^{(r)}\Big\|^2 \tag{116}$$

F.2 Proof of Lemma E.7

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

$$
\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\mathbb{E}\left\|\boldsymbol{w}_j^{(r,c)}-\boldsymbol{w}^{(r)}\right\|^2\\
&\leq 36\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&+10\eta^2L^2\tau\sum_{c=0,r}^{\tau-1}\sum_{c=0,r-1}^{\tau-1}\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{(r-1)}\right\|^2+20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2\\
&\overset{①}{\leq} 36\eta^2\tau^2\sigma^2+\frac{8\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\sum_{c=0}^{\tau-1}\left(\mathbf{g}_j^{(c,0)}-\mathbf{g}^{(0)}\right)\right\|^2\\
&+\Big[10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\sum_{c=0,r-1}^{\tau-1}\left[\left\|\frac{1}{p}\sum_{j=1}^{p}\mathbf{g}_j^{(c,r-1)}\right\|^2\right]\\
&\quad+10\eta^2L^2\tau^4(\eta\gamma)^2(q+1)\frac{\sigma^2}{p}+10\eta^2L^2\tau^3(\eta\gamma)^2(q+1)G_q\Big]+20\eta^2\tau^2\sum_{c=0}^{\tau-1}\left\|\mathbf{g}^{(r)}\right\|^2
\end{aligned}
\tag{117}
$$

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