

# Accelerated Federated Optimization with Quantization

Yejoon Youn<sup>†</sup>      Bhuvesh Kumar<sup>†</sup>      Jacob Abernethy<sup>†</sup>

<sup>†</sup> Georgia Institute of Technology      {yjyou92, bhuvesh, prof}@gatech.edu

## Abstract

*Federated optimization is a new form of distributed training on very large datasets that leverages many devices each containing local data. While decentralized computation can lead to significant speed-ups due to parallelization, some centralization is still required: devices must aggregate their parameter updates through synchronization across the network. The potential for communication bottleneck is significant. The two main methods to tackle this issue are (a) smarter optimization that decreases the frequency of communication rounds and (b) using compression techniques such as quantization and sparsification to reduce the number of bits machines need to transmit. In this paper, we provide a novel algorithm, **Federated optimization algorithm with Acceleration and Quantization (FedAQ)**, with improved theoretical guarantees by combining an accelerated method of federated averaging, reducing the number of training and synchronization steps, with an efficient quantization scheme that significantly reduces communication complexity. We show that in a homogeneous strongly convex setting, FedAQ achieves a linear speedup in the number of workers  $M$  with only  $\tilde{O}(M^{\frac{1}{3}})$  communication rounds, significantly smaller than what is required by other quantization-based federated optimization algorithms. Moreover, we empirically verify that our algorithm performs better than current methods.*

## 1 Introduction

Federated learning (FL) has attracted much attention from both academia and industry due to the increasing demand for large-scale distributed machine learning systems and preserving privacy-sensitive data on local devices such as smartphones and IoT devices. In federated learning, a number of clients collaboratively learn the global objective function by communicating with a central server without sharing any locally stored data in each local device. The research in Federated learning has identified four major challenges: communication efficiency, systems heterogeneity, statistical heterogeneity, and privacy [19]. In this paper, we focus on communication efficiency that is of primary interest in cross-device settings when there is a heavy communication burden with many edge computing devices and limited network bandwidth. Two of the most widely used methods to reduce the communication cost are federated averaging optimization and randomized compression techniques.

In federated averaging (FedAvg) [26], also called *local SGD*, each client locally updates its model with multiple stochastic gradient descent (SGD) steps, and a server aggregates model updates of clients. The server updates its own model parameters by averaging client models and then broadcasts the server parameters to all clients. This enables FL systems to achieve high communication efficiency with infrequent synchronization while showing better performance than distributed large mini-batch SGD [25]. Due to the significant empirical success

---

Copyright 2023 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Table 14: Summary of Results on the Convergence Rate and Communication Required for Linear Speedup.  $M$  is the number of devices,  $T$  is the number of total parallel iterations, and  $K$  is the number of communication rounds,  $q$  is a quantization parameter (Assumption 1),  $d_{\text{quant}}$  is the number of bits used to quantize,  $d_{\text{full}}$  is the number of bits required when there is no quantization ( $d_{\text{full}} \gg d_{\text{quant}}$ ). [43] and FedAQ send two iterates per communication round as other algorithms to achieve acceleration (See line 11 in Algorithm 3), we multiply  $d_{\text{full}}$  and  $d_{\text{quant}}$  by 2 for bits communicated for a linear speedup. The presented results of [9] are newly obtained (section 5.5).

Algorithm	Convergence rate	Communication rounds for $\tilde{\mathcal{O}}(\frac{1}{T})$ convergence with linear speedup	Bits communicated for linear speedup
FedPAQ [27]	$\mathcal{O}(\frac{1+q}{K} + \frac{T}{K^2})$	Not possible	Not possible
FedCOMGATE [9]	$\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1}{TK})$	$\tilde{\mathcal{O}}(\frac{M}{1+q})$	$\tilde{\mathcal{O}}(\frac{M}{1+q}) \cdot d_{\text{quant}}$
FedAC [43]	$\tilde{\mathcal{O}}(\frac{1}{MT} + \frac{1}{TK^3})$	$\tilde{\mathcal{O}}(M^{\frac{1}{3}})$	$\tilde{\mathcal{O}}(M^{\frac{1}{3}}) \cdot 2d_{\text{full}}$
<b>FedAQ</b> (Corollary 23)	$\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$	$\tilde{\mathcal{O}}(M^{\frac{1}{3}})$	$\tilde{\mathcal{O}}(M^{\frac{1}{3}}) \cdot 2d_{\text{quant}}$

of FedAvg, researchers have proposed an interesting theoretical question: To what extent can we minimize the number of synchronizations in order to both guarantee convergence and achieve linear speedup in the number of workers  $M$ <sup>1</sup>? For the strongly-convex and homogeneous settings, [15] was able to achieve a linear speedup in  $M$  with  $\tilde{\mathcal{O}}(M)$  communication rounds, which is the state-of-the-art result for FedAvg convergence analysis. However, even with this progress on theoretical guarantees of FedAvg, it remains unclear whether further improvements on convergence time and communication efficiency can be achieved.

Applying acceleration methods to FL has led to improved convergence, with [43] providing a faster version of FedAvg with provably stronger bounds. For the strongly-convex and homogeneous setting, their algorithm achieves a linear speedup in  $M$  with only  $\tilde{\mathcal{O}}(M^{\frac{1}{3}})$  communication rounds. Hence, the accelerated version of federated averaging requires a much smaller number of communication rounds than FedAvg to achieve the same accuracy. At present, this remains the best result for strongly-convex and homogeneous local data distribution settings. In addition to reducing the required number of communication rounds, another powerful way to build communication-efficient FL systems is to reduce the number of bits that need to be transmitted at each synchronization. [27, 9] have shown that such compression techniques, which include *quantization*, reduce communication costs and guarantee convergence (See Table 14).

In this work, we provide a novel algorithm, **F**ederated optimization algorithm with **A**cceleration and **Q**uantization (FedAQ), to solve the severe communication bottleneck problem in FL systems. FedAQ is the first federated optimization algorithm that successfully incorporates multiple local update schemes, acceleration, and quantization for master-worker topology. Although these three key desiderata of Federal Learning systems have individually been shown to build communication-efficient FL systems, it is not obvious if or how acceleration techniques can lead to faster convergence even for quantization-based methods. We answer this question by showing that FedAQ converges for strongly-convex and homogeneous local data distribution settings without any additional strong assumptions.

Let  $T$  be the number of total parallel iterations,  $K$  be the number of total communication rounds. We compare our results to previous methods in Table 14, and highlight the following contributions:

1. FedAQ has a convergence rate of  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$  which is better than the  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1}{TK})$  convergence of [9], the state of the art in quantization based methods. Here  $q$  is a parameter that measures the effectiveness of the quantization scheme (see Assumption 1). This allows FedAQ to obtain linear speedup with only  $\tilde{\mathcal{O}}(M^{\frac{1}{3}})$  communication rounds whereas [9] requires  $\tilde{\mathcal{O}}(\frac{M}{1+q})$  rounds. The faster convergence in number of

<sup>1</sup>Linear speedup in the number of workers is a desirable property in parallel computing which implies that the task takes half as much time if the number of workers are doubled.

communication rounds also implies that FedAQ can achieve better convergence than [9] by using many fewer communication rounds. Thus, although FedAQ sends two iterates in each communication round, that is the bits communicated in each round are twice many compared to [9] for the same level of quantization, FedAQ requires much smaller total communication costs due to the large reduction in synchronization rounds.

2. When comparing FedAQ to Accelerated Federated learning, we observe that FedAQ has similar convergence and requires the same number of communication rounds as [43]. In each communication round of [43], every client sends the complete iterates to the server without any quantization. To effectively obtain a convergence rate of  $\tilde{\mathcal{O}}(\frac{1}{MT})$ , it needs to send each value with a precision of  $\tilde{\mathcal{O}}(\frac{1}{MT})$ , requiring  $d_{\text{full}} = \mathcal{O}(\log(MT))$  bits. In comparison, if we use the low precision quantizer (See section 3 Example 1) given by [1], FedAQ needs to send only  $d_{\text{quant}} = \mathcal{O}(\log \frac{1}{q})$  bits <sup>2</sup> for each value. Since  $q$  is a constant,  $d_{\text{quant}} \ll d_{\text{full}}$ . The extra  $1 + q$  term in the convergence for FedAQ can be offset by scaling the number of local updates by  $1 + q$ , which is cheaper than expensive data communication. Thus, FedAQ obtains the same convergence as [43] using as many communication rounds but by sending many fewer bits per round.

Finally, we empirically verify that our algorithm exhibits better performance than baselines, FedPAQ [27], FedCOMGATE [9], FedAC [43], and FedAvg [26] on classical vision datasets such as MNIST [18] and CIFAR-10 [17].

## 2 Related Works

The first guarantee for FedAvg, showing that it converges at the same rate as mini-batch SGD in strongly convex scenarios, was shown by [30] in the IID setting. The further convergence analysis of FedAvg for non-convex functions was laid out in a number of published works [36, 8, 42]. Followup work has managed to remove unnecessary assumptions, such as uniformly bounded gradients, to achieve better convergence rates [36, 31, 7, 15, 40]. Moreover, [20, 10, 21, 15, 14] define scenarios that depart from the IID framework, analyzing the convergence of FedAvg and its variants in settings with heterogeneous data distributions.

Reducing the transmitted bits between a server and clients through compression techniques is pivotal to saving communication costs in federated learning. This motivates researchers to develop various compression techniques such as sparsification and quantization without significantly sacrificing accuracy [16, 1, 32, 39, 4, 34, 33, 11, 3, 28]. [27] shows near-optimal theoretical guarantees of the first federated optimization algorithm that incorporates federated averaging, partial node participation, and quantization in homogeneous local data distribution settings. [9] further provide improved convergence rates for both homogeneous and heterogeneous settings.

We can achieve better communication efficiency by applying acceleration methods into client updates. [43] has proposed the first provable acceleration of FedAvg that achieves a linear speedup with the fewest communication rounds. Several other works aim to achieve communication efficiency by using momentum or adaptive optimizers [41, 13, 37]. It is important to note that our work is not the first to combine acceleration and quantization. [23, 24], for example, propose compressed and accelerated distributed optimization methods that are neither stochastic nor FedAvg variants. [29] proposes communication efficient momentum SGD for decentralized optimization. [22, 38] show that distributed and federated versions of adaptive optimizers along with gradient compression can lead to similar convergence rates as their non-compressed counterparts. But these works do not achieve the core result of the present paper, which is the reduced communication complexity via a faster convergence rate and a linear speedup with the small number of communication rounds. To the best of our knowledge, FedAQ is the first accelerated version of federated averaging for master-worker topology that successfully integrates a quantization scheme and provides rigorous convergence guarantees.

---

<sup>2</sup>More details on this are discussed in section 5.5.3

### 3 Problem Setup

In this paper, we build our algorithm based on federated learning with captain-worker topology where  $M$  local devices contain their own local data, and a server aggregates local parameter updates without sharing any data during synchronization rounds. Since we focus on homogeneous local data distribution settings for the convergence analysis of our algorithm, we define the distributed stochastic optimization problem as below.

$$\min_{w \in \mathbb{R}^d} F(w) := \mathbb{E}_{z \sim \mathcal{D}}[f(w; z)]$$

In our convergence analysis, we assume  $F$  is strongly-convex. Each client can access  $F$  at  $w$  via oracle  $\nabla f(w; z)$  because all clients have the same loss function  $f$ . Also, every local device has the same local data distribution  $\mathcal{D}$ . Moreover, we use the full participation of nodes for local updates and synchronizations.

#### 3.1 Assumptions

Let us clarify assumptions on the unbiased quantizer  $Q$ , the global objective function  $F$ , and the unbiased gradient estimator  $\nabla f$ .

**Assumption 1:** The variance of the unbiased quantizer  $Q$  is bounded by the squared of  $l_2$ -norm of its argument, i.e.,  $\mathbb{E}[Q(x)|x] = x$ ,  $\mathbb{E}[\|Q(x) - x\|^2|x] \leq q\|x\|^2$ .

For example, a well-known randomized quantizer which satisfies Assumption 1 is low-precision quantizer in [1].

**Example 1.** (Low-precision quantizer) Given  $x \in \mathbb{R}^d$ , the quantizer  $Q : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is defined by

$$Q_i(x) = \text{sign}(x_i) \cdot \|x\| \cdot \xi_i(x, s), \quad i \in [d]$$

$\xi_i$  is defined as below.

$$\xi_i(x, s) = \begin{cases} \frac{l+1}{s}, & \text{with probability } \frac{|x_i|}{\|x\|}s - l \\ \frac{l}{s}, & \text{o/w} \end{cases}$$

$s$  is the number of quantization levels.  $l \in [0, s)$  is an integer which satisfies  $\frac{|x_i|}{\|x\|} \in [\frac{l}{s}, \frac{l+1}{s})$ .

**Assumption 2:**  $F$  is  $\mu$ -strongly convex, i.e.,  $F(w_1) \geq F(w_2) + \langle \nabla F(w_2), w_1 - w_2 \rangle + \frac{1}{2}\mu\|w_1 - w_2\|^2$  for any  $w_1, w_2 \in \mathbb{R}^d$ .

**Assumption 3:**  $F$  is  $L$ -smooth, i.e.,  $F(w_1) \leq F(w_2) + \langle \nabla F(w_2), w_1 - w_2 \rangle + \frac{1}{2}L\|w_1 - w_2\|^2$  for any  $w_1, w_2 \in \mathbb{R}^d$ .

**Assumption 4:**  $\nabla f(w; \xi)$  is unbiased and variance bounded, i.e.,  $\mathbb{E}_\xi[\nabla f(w; \xi)] = \nabla F(w)$ ,  $\mathbb{E}_\xi[\|\nabla f(w; \xi) - \nabla F(w)\|^2] \leq \sigma^2$  for any  $w \in \mathbb{R}^d$ .

#### 3.2 Notation

We use  $\tau, K$  to respectively denote the number of local updates and total communication rounds, which means the total number of iterations  $T$  at each node satisfies  $T = K\tau$ . Since we consider a strongly-convex case, we can find the optimal point  $w^*$  and denote the optimal function value as  $F^* := F(w^*)$ . The local parameter  $w_{k,t}^m$  indicates the parameter of the  $m$ -th local model after  $k$ th synchronization followed by  $t$  local SGD updates. There are other types of parameters such as  $w_{k,t}^{\text{ag},m}$  and  $w_{k,t}^{\text{md},m}$ , and we obtain two types of parameters  $w_k$  and  $w_k^{\text{ag}}$  in the server side after  $k$ th synchronization. More details on these parameters will be discussed in the next section.

## 4 FedAQ Algorithm

We propose a novel communication efficient algorithm that combines an accelerated variant of federated averaging and an efficient quantization scheme. Our FedAQ algorithm has two main parts: (1) multiple accelerated local updates and (2) communication with quantization. Both components contribute to achieving better communication efficiency than other previous federated algorithms. The entire process is summarized in Algorithm 3.

### 4.1 Multiple Accelerated Local Updates

The FedAvg algorithm, proposed by [26], is widely used for federated learning to improve communication efficiency by reducing communication rounds with multiple local SGD updates. [43] provide FedAC that replaces the stochastic gradient updates of FedAvg by accelerated version of SGD by [6] resulting in a linear speedup in  $M$  with fewer communication rounds than FedAvg.

Thus, we apply the FedAC scheme to multiple updates of each local model. Since previous quantization-based federated optimization algorithms are FedAvg variants with no acceleration, the accelerated method enables our algorithm to gain better communication efficiency than others.

As you can see in Algorithm 3, we need two more local parameters  $w_{k,t}^{\text{ag},m}$  and  $w_{k,t}^{\text{md},m}$  for acceleration in addition to the main local parameter  $w_{k,t}^m$ .  $w_{k,t}^{\text{ag},m}$  aggregates the past iterates, and the gradients are queried at the auxiliary parameter  $w_{k,t}^{\text{md},m}$ . While typical FL algorithms without acceleration only have a learning rate  $\eta$  as their hyperparameter, the general acceleration scheme makes our algorithm flexible due to four hyperparameters  $\alpha, \beta, \eta, \gamma$ .  $\alpha, \beta$  are hyperparameters related to coupling coefficients, and  $\eta, \gamma$  stand for learning rates respectively for  $w_{k,t}^{\text{ag},m}, w_{k,t}^m$ . The flexibility of hyperparameters enables the fast convergence speed of FedAQ, but naively chosen hyperparameters also cause unstable training of FedAQ. We discuss the exact choice of hyperparameters in section 5. Unlike FedAC, that requires each client to communicate the exact iterates to the server with high precision, we discuss in the following subsection how FedAQ incorporates quantization techniques to reduce communication cost.

### 4.2 Communication with Quantization

In cross-device federated learning, a large amount of communicated messages from a number of devices and the limited communication bandwidth can lead to severe communication bottlenecks. Therefore, in this scenario, an efficient quantization scheme can significantly reduce the size of communicated messages and make communication between local devices and a server faster. We apply the same unbiased quantizer used in [9] that satisfies Assumption 1.

In contrast with other quantization-based federated optimization algorithms [27, 9], the algorithmic novelty of FedAQ is based on applying quantization to two model parameter updates, which is required in order to simultaneously reduce the frequency of communication and the volume of communicated bits. To the best of our knowledge, this is the first quantization-based method that achieves the accelerated rate with the dramatic reduction in communication cost. To be specific on the communication process, after each client  $m$  obtains  $w_{k,\tau}^m, w_{k,\tau}^{\text{ag},m}$  through  $\tau$  accelerated local iterations, each client quantizes the difference between  $w_{k,\tau}^m, w_{k,\tau}^{\text{ag},m}$  and the most recent server models  $w_k, w_k^{\text{ag}}$ . Then, a server aggregates  $Q(w_{k,\tau}^m - w_k), Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})$  from all clients. After dequantizing those messages, the server obtains the following new models  $w_{k+1}, w_{k+1}^{\text{ag}}$  and broadcasts them back to each client.

---

**Algorithm 3** Federated Accelerated SGD with Quantization (FedAQ)

---

```

1: Input:  $\alpha, \beta, \eta, \gamma$ , initial vector  $w_0 = w_{0,0}^{\text{ag},m} = w_{0,0}^m$  for all devices  $m \in [M]$ 
2: for  $k = 0, \dots, K - 1$  do
3:   for each client  $m$  in parallel do
4:      $w_{k,0}^m \leftarrow w_k, w_{k,0}^{\text{ag},m} \leftarrow w_k^{\text{ag}}$ 
5:     for  $t = 0, \dots, \tau - 1$  do
6:        $w_{k,t}^{\text{md},m} \leftarrow \beta^{-1}w_{k,t}^m + (1 - \beta^{-1})w_{k,t}^{\text{ag},m}$ 
7:        $g_{k,t}^m \leftarrow \nabla f(w_{k,t}^{\text{md},m}, \xi_{k,t}^m)$ 
8:        $w_{k,t+1}^{\text{ag},m} \leftarrow w_{k,t}^{\text{md},m} - \eta g_{k,t}^m$ 
9:        $w_{k,t+1}^m \leftarrow (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma g_{k,t}^m$ 
10:      send  $Q(w_{k,\tau}^m - w_k), Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})$ 
11:  server finds  $w_{k+1} \leftarrow w_k + \frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^m - w_k), w_{k+1}^{\text{ag}} \leftarrow w_k^{\text{ag}} + \frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})$ 

```

---

## 5 Convergence Analysis

The rigorous theoretical guarantees of reducing communication complexity under strongly-convex and homogeneous assumptions should come first to ensure the significance of FedAQ as one of the standards of communication-efficient federated optimization algorithms. Proving convergence guarantees of FedAQ even under these assumptions requires careful consideration of the approximation error induced by the quantization scheme combined with the convergence analysis of acceleration based methods. To recall, in FedAQ the server aggregates two quantized local updates  $Q(w_{k,\tau}^m - w_k), Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})$  from all clients (See line 11 in Algorithm 3) in each round. If we simply try to generalize the convergence guarantee of FedAC to incorporate the quantization variance costs, the proof techniques from earlier quantization-based methods cannot be directly applied, as we now have two additional quantization error terms that contribute to the overall cost. A significant amount of additional effort is required in order to account for this new quantization error.

In this section, we first define two condition sets of hyperparameters used for the convergence analysis of FedAQ. Then, we provide the proof sketch of FedAQ under one such condition set that leads to the better convergence rate  $\tilde{O}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$ . The full proofs of lemmas, theorems, and corollaries under both condition sets are elaborated in section 5.3 and section 5.4. Finally, we discuss how we obtain the new convergence rate for [9] and look into more theoretical details on contribution 2 in Introduction.

### 5.1 Two Parameter Condition Sets

We carefully determine two parameter condition sets that theoretically ensure the convergence guarantees. The first one is

$$\eta, \gamma \in \left(0, \frac{1}{L}\right], \gamma = \max\left(\sqrt{\frac{\eta}{\mu\tau}}, \eta\right), \alpha = \frac{1}{\gamma\mu}, \beta = \alpha + 1 \quad (34)$$

We add one more condition  $\gamma \in (0, \frac{1}{L}]$  to the FedAC-I condition [43] and create our parameter condition set (34). The second one is

$$\eta, \gamma \in \left(0, \frac{1}{L}\right], \gamma = \max\left(\sqrt{\frac{\eta}{\mu\tau}}, \eta\right), \alpha = \frac{3}{2\gamma\mu} - \frac{1}{2}, \beta = \frac{2\alpha^2 - 1}{\alpha - 1}, \gamma\mu \leq \frac{3}{4} \quad (35)$$

We add two more conditions  $\gamma \in (0, \frac{1}{L}]$  and  $\gamma\mu \leq \frac{3}{4}$  to the FedAC-II condition to build our parameter condition set (35). Even though quantization adds complexity to the algorithm, these weak assumptions are the only

additional requirements for showing the convergence of FedAQ. Moreover, although the better convergence rate  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$  is obtained from the condition set (35), we also analyze the convergence of FedAQ under the condition set (34) because this set empirically leads to more stable training and better performance in experiments than the condition set (35) (See Strongly convex case in section 6.2.1). The intuition of the less stable training of FedAQ under the condition set (35) comes from larger  $\alpha, \beta$  than those of the condition set (34). If  $\alpha, \beta$  are too large,  $\alpha^{-1}, \beta^{-1}$  in Algorithm 3 cannot be used as proper coupling coefficients for local parameters  $w_{k,t}^m, w_{k,t}^{\text{ag},m}, w_{k,t}^{\text{md},m}$ . This results in aggressive updates and less stable training behavior.

## 5.2 Proof Sketch of FedAQ Under Condition Set (35)

The decentralized potential  $\Phi_{k,t}$  [43] is used for our convergence analysis. People commonly use this potential for acceleration analysis [2].

$$\Phi_{k,t} = F(\bar{w}_{k,t}^{\text{ag}}) - F^* + \frac{1}{6}\mu\|\bar{w}_{k,t} - w^*\|^2$$

$\bar{w}_{k,t}$  and  $\bar{w}_{k,t}^{\text{ag}}$  is respectively the average of  $w_{k,t}^m$  and  $w_{k,t}^{\text{ag},m}$  for all  $m$ . Here, we additionally define  $\Phi_k$  as below.

$$\Phi_k := \Phi_{k,0} = F(w_k^{\text{ag}}) - F^* + \frac{1}{6}\mu\|w_k - w^*\|^2$$

Since  $w_k$  and  $w_k^{\text{ag}}$  are parameters obtained after  $k$ th synchronization in a server side,  $\Phi_k$  can be considered as the potential of server models.  $\Phi_k$  is essential to show the convergence of FedAQ because there is the computation of the quantizer between  $\Phi_{k-1,\tau}$  and  $\Phi_{k,0}$ . Thus, we should not naively track  $\Phi_{k,t}$  but track  $\Phi_k$  for our analysis. Obtaining  $\Phi_k \leq \epsilon$  would imply that  $F(w_k^{\text{ag}}) - F^* \leq \epsilon$  and since  $F^* \leq F(w_k^{\text{ag}})$ , it would also imply that  $\|w_k - w^*\|^2 = O(\epsilon)$ , thus obtaining convergence in terms of both the objective value and the iterate.

Our goal is to show the convergence of FedAQ and derive the simplified convergence rate so that we can get the number of communication rounds to achieve a linear speedup in  $M$ . As the first step to show this, we prove Lemma 5 which represents the relationship between two consecutive server potential functions  $\Phi_k$  and  $\Phi_{k+1}$ . The quantization scheme amplifies the instability to the convergence of FedAQ in addition to the effect of acceleration. Despite this challenge, we derive Lemma 5 with the help of subtle Propositions (See section 5.4.1).

**Lemma 5:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 1, 2, 3, 4, then for  $\alpha = \frac{3}{2\gamma\mu} - \frac{1}{2}, \beta = \frac{2\alpha^2 - 1}{\alpha - 1}, \gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}], \eta, \gamma \in (0, \frac{1}{L}], \gamma\mu \leq \frac{3}{4}, \tau \geq 2$ , FedAQ yields

$$\begin{aligned} \mathbb{E}[\Phi_{k+1}] &\leq D(\gamma, \tau)\mathbb{E}[\Phi_k] + \left(\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6}\right)\frac{\tau\sigma^2}{M} + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ &\quad + \underbrace{\frac{q}{M}\left(\frac{\gamma^2 \mu}{3} + \eta^2 L\right)\tau\sigma^2 + \frac{q}{2M}\left((\gamma - \eta)^2 \gamma^2 \mu^2 \left(\frac{\mu}{3} + \frac{L}{4}\right) + \gamma^4 \left(\frac{\mu}{3} + L\right)^2 L\right)\tau^3 \sigma^2}_{\text{additional terms due to quantization}} \end{aligned}$$

Where  $D(\gamma, \tau)$  is defined as

$$D(\gamma, \tau) = \left(1 - \frac{1}{3}\gamma\mu\right)\tau + \underbrace{\frac{q}{M}\left(\gamma^2 \mu \left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2 L \left(\frac{\mu}{3} + L\right)\right)\tau^2}_{\text{additional terms due to quantization}}$$

We get the inequality between  $\Phi_k$  and  $\Phi_{k+1}$  by finding the upper bounds of error terms due to multiple( $\tau$ ) local steps and the quantization step. The upper bound of the error caused by multiple local steps is obtained

with the help of the analysis in [43] (See Proposition 19). Also, we get the tight upper bound of the error due to quantization with our new proof techniques (See Proposition 20, 21, 22). The key challenge in bounding the quantization error terms comes from representing the upper bound of variances of the quantizer  $Q$  on two local updates  $w_{k,\tau}^m - w_k, w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}$  in the form of a server potential  $\Phi_k$ . Some terms in Lemma 5 are similar to those in Lemma C.2 of the FedAC paper [43], but our lemma contains additional terms that emerge from the quantization scheme.

For the next step, by telescoping Lemma 5, we obtain the main theoretical result Theorem 6. Theorem 6 represents how  $\Phi_K$  decreases from the initial potential  $\Phi_0$  as a communication round  $K$  increases. Since we aim to telescope Lemma 5,  $D(\gamma, \tau)$  should be smaller than 1. Specifically, we show  $D(\gamma, \tau) \leq 1 - \frac{1}{6}\gamma\mu\tau$  with condition (36) (See section 5.4.2). That's why Theorem 6 requires the learning rate  $\gamma$  to satisfy the certain condition (36).

**Theorem 6:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 1, 2, 3, 4, then for the parameter condition set (35),  $\tau \geq 2$ , if the learning rate  $\gamma$  satisfies

$$\left( \frac{1}{9}\mu^2 + \frac{q}{M} \left( \mu \left( \frac{8}{3}\mu + 2L \right) + 2L \left( \frac{\mu}{3} + L \right) \right) \right) \gamma\tau \leq \frac{1}{6}\mu \quad (36)$$

FedAQ yields

$$\begin{aligned} \mathbb{E}[\Phi_K] \leq & \exp\left(-\frac{1}{6} \max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}})K\tau\right)\Phi_0 + \frac{2(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}M\tau^{\frac{1}{2}}} + \frac{8(q+25)\eta^2L^2\tau\sigma^2}{\mu} \\ & + \frac{3q\left(\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + L\left(\frac{\mu}{3} + L\right)^2\right)\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{5}{2}}M} + \frac{3qL\left(\frac{\mu}{3} + L\right)^2\eta^3\tau^2\sigma^2}{\mu M} \end{aligned}$$

We get the convergence rate of FedAQ with respect to  $\eta$  under the condition set (35). The final step is to tune  $\eta$  appropriately and obtain a more intuitive form of convergence rate that we can easily analyze a linear speedup in  $M$ . The exact form of this can be found in Corollary 23. Here, we introduce the simplified form of Corollary 23.

**Corollary 7:** (Simplified form of Corollary 23) Note that  $T = K\tau$ . For  $\eta = \min\left(\frac{1}{L}, \tilde{\Theta}\left(\frac{\tau}{\mu T^2}\right)\right)$ , FedAQ yields

$$\mathbb{E}[\Phi_K] \leq \min\left(\exp\left(-\frac{\mu T}{6L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}}T}{6L^{\frac{1}{2}}\tau^{\frac{1}{2}}}\right)\right)\Phi_0 + \underbrace{\tilde{\mathcal{O}}\left(\frac{(1+q)\sigma^2}{\mu MT}\right)}_{\text{I}} + \underbrace{\tilde{\mathcal{O}}\left(\frac{(1+q)L^2\tau^3\sigma^2}{\mu^3T^4}\right)}_{\text{II}} + \underbrace{\tilde{\mathcal{O}}\left(\frac{qL^3\tau^2\sigma^2}{\mu^4MT^3}\right)}_{\text{III}}$$

The convergence rate of FedAQ under the condition set (34) is obtained in a similar way. The convergence analysis under the condition set (34) is elaborated as Lemma 10, Theorem 16, and Corollary 17 in section 5.3.

**Remark 8:** The above convergence rate is worse than the convergence rate of FedAC-II according to Theorem C.13 in [43] because there are additive terms related to the quantization noise  $q$  in our case. Let's figure out the dominant terms with  $\tilde{\mathcal{O}}$  notation from the above convergence rate. Here, we replace  $\tau$  with  $\frac{T}{K}$ . At first, we can ignore the first term because it decreases exponentially. The second term I would be  $\tilde{\mathcal{O}}\left(\frac{1+q}{MT}\right)$ . Then, the third term II becomes  $\tilde{\mathcal{O}}\left(\frac{(1+q)\tau^3}{T^4}\right) = \tilde{\mathcal{O}}\left(\frac{1+q}{TK^3}\right)$ . Finally, the last term III turns into  $\tilde{\mathcal{O}}\left(\frac{q\tau^2}{MT^3}\right) = \tilde{\mathcal{O}}\left(\frac{q}{MTK^2}\right)$ . Thus, the overall convergence rate of FedAQ under the condition set (35) would be  $\tilde{\mathcal{O}}\left(\frac{1+q}{MT} + \frac{1+q}{TK^3}\right)$ . Similarly, we obtain the simplified convergence rate of FedAQ under the condition set (34) from three terms (14), (15), (16) of Corollary 17. In this case, the convergence rate of FedAQ is  $\tilde{\mathcal{O}}\left(\frac{1+q}{MT} + \frac{1}{TK^2}\right)$ , and the required number of communication rounds to achieve a linear speedup in  $M$  is  $\tilde{\mathcal{O}}\left(\left(\frac{M}{1+q}\right)^{\frac{1}{2}}\right)$ .



**Remark 9:** As we mention above, FedAQ converges at rate  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$ , which is better than the convergence rate of [9]  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1}{TK})$ . To our knowledge, [9] obtain the best convergence rate among previous quantization-based federated optimization algorithms. Actually, in the strongly-convex and homogeneous case, [9] provide different convergence rate  $\mathcal{O}(\frac{1}{\gamma^2\tau} + \frac{(q+1)}{(\frac{q}{M}+1)\tau M}) = \mathcal{O}(\frac{K}{\gamma^2T} + \frac{(q+1)K}{(\frac{q}{M}+1)TM})$ , where  $\gamma$  is a learning rate for the server updates. They achieve this convergence rate by tuning  $\eta = \frac{1}{2L(\frac{q}{M}+1)\tau\gamma}$ . However, we cannot say this algorithm achieves a linear speedup in this scenario. That's why we provide a new convergence rate  $\tilde{\mathcal{O}}(\frac{1+q}{MT} + \frac{1}{TK})$  for [9] by tuning  $\eta$  in a different way. This new  $\eta$  makes this algorithm achieve a linear speedup. Why the original  $\eta$  cannot achieve a linear speedup and how we get new  $\eta$  can be found in section 5.5.

### 5.3 Proof Details for FedAQ under Condition Set (34)

Before diving into proof details, we define  $\bar{w}_{k,\tau}, \bar{w}_{k,\tau}^{\text{ag}}, \Psi_{k,t}^m, \Psi_{k,t}, \Psi_k, A_{k,t}^m$  as below.

$$\begin{aligned}\bar{w}_{k,\tau} &= \frac{1}{M} \sum_{m=1}^M w_{k,\tau}^m \\ \bar{w}_{k,\tau}^{\text{ag}} &= \frac{1}{M} \sum_{m=1}^M w_{k,\tau}^{\text{ag},m} \\ \Psi_{k,t}^m &= F(w_{k,t}^{\text{ag},m}) - F^* + \frac{1}{2}\mu \|w_{k,t}^m - w^*\|^2 \\ \Psi_{k,t} &= \frac{1}{M} \sum_{m=1}^M F(w_{k,t}^{\text{ag},m}) - F^* + \frac{1}{2}\mu \|\bar{w}_{k,t} - w^*\|^2 \\ \Psi_k &:= \Psi_{k,0} = F(w_k^{\text{ag}}) - F^* + \frac{1}{2}\mu \|w_k - w^*\|^2 \\ A_{k,t}^m &= \frac{\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} \|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + \gamma^2(\mu + L) \frac{2L}{1 + \gamma\mu} \Psi_{k,t}^m\end{aligned}$$

The above notations are essential to our convergence analysis. Intuitively, if the FedAQ algorithm converges to the optimal point,  $\bar{w}_{k,\tau}, \bar{w}_{k,\tau}^{\text{ag}}$  become  $w^*$ , and  $\Psi_{k,t}^m, \Psi_{k,t}, \Psi_k, A_{k,t}^m$  become 0. In order to denote the  $\sigma$ -algebra generated by  $\{w_{k',t'}^m, w_{k',t'}^{\text{ag},m}\}_{(k' < k) \text{ or } (k'=k, t' \leq t), m \in [M]}$ , we use  $\mathcal{F}_{k,t}$ .

#### 5.3.1 Proof of Lemma 10

**Lemma 10:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 1, 2, 3, 4, then for  $\alpha = \frac{1}{\gamma\mu}, \beta = \alpha + 1, \gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}], \eta, \gamma \in (0, \frac{1}{L}], \tau \geq 2$ , FedAQ yields

$$\begin{aligned}\mathbb{E}[\Psi_{k+1}] &\leq C(\gamma, \tau)\mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2L + \frac{\gamma^2\mu}{M})\tau\sigma^2 \\ &\quad + \gamma\mu L\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \|\frac{1}{1 + \gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})\|] \\ &\quad + \underbrace{\frac{q}{M}(\gamma^2\mu + \eta^2L)\tau\sigma^2 + \frac{q}{2M} \left( \frac{(\gamma - \eta)^2\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2L}{1 + \gamma\mu} \right)}_{\text{Additional terms due to quantization}}\tau^3\sigma^2\end{aligned}$$

Where  $C(\gamma, \tau)$  is defined as

$$C(\gamma, \tau) = (1 - \gamma\mu)^\tau + \underbrace{\frac{q}{M} \left( \frac{4\gamma^2\mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right)}_{\text{Additional terms due to quantization}} \tau^2$$

In this section, we first introduce five crucial Propositions for proving Lemma 10. Then, we prove Lemma 10 by using Propositions in the last part of this section.

**Proposition 11:** Let Assumption 1 hold and consider any  $k$  synchronization round. Then, we can decompose the expectation as follows:

$$\begin{aligned} \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] + \mathbb{E}[\|\bar{w}_{k,\tau} - w^*\|^2] \\ \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F^*] &= \mathbb{E}[F(w_{k+1}^{\text{ag}}) - \frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m})] + \mathbb{E}[\frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m}) - F^*] \end{aligned}$$

Proof of Proposition 11 The second equality is trivial. Let's focus on the first equality. By Assumption 1, the quantizer  $Q$  is unbiased and we get,

$$\mathbb{E}_Q[w_{k+1}] = w_k + \frac{1}{M} \sum_{m=1}^M \mathbb{E}_Q Q(w_{k,\tau}^m - w_k) = \frac{1}{M} \sum_{m=1}^M w_{k,\tau}^m = \bar{w}_{k,\tau}$$

Thus, we finally obtain

$$\begin{aligned} \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau} + \bar{w}_{k,\tau} - w^*\|^2] \\ &= \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] + \mathbb{E}[\|\bar{w}_{k,\tau} - w^*\|^2] \end{aligned}$$

**Proposition 12:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{1}{\gamma\mu}$ ,  $\beta = \alpha + 1$ ,  $\gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}]$ ,  $\eta \in (0, \frac{1}{L}]$ , FedAQ yields

$$\begin{aligned} \mathbb{E}[\Psi_{k,\tau}] &\leq (1 - \gamma\mu)^\tau \mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2 L + \frac{\gamma^2\mu}{M})\tau\sigma^2 + \gamma\mu L\tau \\ &\cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \frac{1}{1 + \gamma\mu} (\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu} (\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})] \end{aligned}$$

Proof of Proposition 12 We refer to the proof of Lemma B.2 in [43]. There is no quantization between  $\Psi_{k,\tau}$  and  $\Psi_k$ . Thus, we can directly apply useful inequalities in the proof of Lemma B.2 in [43] to our proof. Then, we obtain

$$\begin{aligned} \mathbb{E}[\Psi_{k,t+1} | \mathcal{F}_{k,t}] &\leq (1 - \gamma\mu)\Psi_{k,t} + \frac{1}{2}(\eta^2 L + \frac{\gamma^2\mu}{M})\sigma^2 + \gamma\mu L \\ &\cdot \frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \frac{1}{1 + \gamma\mu} (\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu} (\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m}) \end{aligned}$$

From the above relationship between  $\Psi_{k,t+1}$  and  $\Psi_{k,t}$ , we get

$$\begin{aligned}
\mathbb{E}[\Psi_{k,\tau}] &\leq (1 - \gamma\mu)^\tau \mathbb{E}[\Psi_k] + \left( \sum_{t=0}^{\tau-1} (1 - \gamma\mu)^t \right) \frac{1}{2} (\eta^2 L + \frac{\gamma^2 \mu}{M}) \sigma^2 + \gamma\mu L \cdot \sum_{t=0}^{\tau-1} \left\{ (1 - \gamma\mu)^{\tau-t-1} \right. \\
&\quad \left. \mathbb{E} \left[ \frac{1}{M} \sum_{m=1}^M \left\| \bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m} \right\| \left\| \frac{1}{1 + \gamma\mu} (\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu} (\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m}) \right\| \right] \right\} \\
&\leq (1 - \gamma\mu)^\tau \mathbb{E}[\Psi_k] + \frac{1}{2} (\eta^2 L + \frac{\gamma^2 \mu}{M}) \tau \sigma^2 + \gamma\mu L \tau \\
&\quad \cdot \max_{0 \leq t < \tau} \mathbb{E} \left[ \frac{1}{M} \sum_{m=1}^M \left\| \bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m} \right\| \left\| \frac{1}{1 + \gamma\mu} (\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu} (\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m}) \right\| \right]
\end{aligned}$$

**Proposition 13:** Let Assumption 1 hold. Then, we have

$$\begin{aligned}
\mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] &\leq \frac{q}{M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] \\
\mathbb{E}[F(w_{k+1}^{\text{ag}}) - \frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m})] &\leq \frac{qL}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2]
\end{aligned}$$

Proof of Proposition 13 First, let's consider the first inequality. According to Assumption 1, we get

$$\begin{aligned}
\mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] &= \mathbb{E}[\|w_k + \frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^m - w_k) - \frac{1}{M} \sum_{m=1}^M w_{k,\tau}^m\|^2] \\
&= \mathbb{E}[\|\frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^m - w_k) - (w_{k,\tau}^m - w_k)\|^2] \\
&= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E}[\|Q(w_{k,\tau}^m - w_k) - (w_{k,\tau}^m - w_k)\|^2] \leq \frac{q}{M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2]
\end{aligned}$$

The third equality comes from the unbiasedness of  $Q$ , and the last inequality stems from the variance assumption of  $Q$ . Similarly, we obtain

$$\begin{aligned}
\mathbb{E}[F(w_{k+1}^{\text{ag}}) - \frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m})] &= \mathbb{E}[F(w_k^{\text{ag}} + \frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})) - \frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m})] \\
&= \mathbb{E}[\frac{1}{M} \sum_{m=1}^M F(w_k^{\text{ag}} + \frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})) - F(w_{k,\tau}^{\text{ag},m})] \\
&\leq \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \langle \nabla F(w_{k,\tau}^{\text{ag},m}), \frac{1}{M} \sum_{m=1}^M (Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})) \rangle + \frac{L}{2} \|\frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2] \\
&= \frac{L}{2} \mathbb{E}[\|\frac{1}{M} \sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2] \\
&= \frac{L}{2M^2} \sum_{m=1}^M \mathbb{E}[\|Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2] \\
&\leq \frac{qL}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2]
\end{aligned}$$

**Proposition 14:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{1}{\gamma\mu}, \beta = \alpha + 1, \gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}], \eta, \gamma \in (0, \frac{1}{L}]$ , we get

$$\mathbb{E}[A_{k,t}^m] \leq \mathbb{E}[A_{k,0}^m] + \left( \frac{(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \frac{\gamma^2(\mu + L)^2 L}{\mu^2} \right) \cdot \left( 1 - (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t \right) \sigma^2$$

Proof of Proposition 14 From the notation mentioned in the beginning of section 5.3,

$$\mathbb{E}[A_{k,t+1}^m | \mathcal{F}_{k,t}] = \frac{\gamma^2 \mu^2 (\mu + L)}{(1 + \gamma\mu)^2} \mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] + \gamma^2 (\mu + L) \frac{2L}{1 + \gamma\mu} \mathbb{E}[\Psi_{k,t+1}^m | \mathcal{F}_{k,t}] \quad (37)$$

Thus, let's sequentially compute  $\mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}]$  and  $\mathbb{E}[\Psi_{k,t+1}^m | \mathcal{F}_{k,t}]$ .

$$\begin{aligned}
\mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] &= \mathbb{E}[\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma g_{k,t}^m - w_{k,t}^{\text{md},m} + \eta g_{k,t}^m\|^2 | \mathcal{F}_{k,t}] \\
&= \mathbb{E}[\|(1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}) - (\gamma - \eta)g_{k,t}^m\|^2 | \mathcal{F}_{k,t}] \quad (\leftarrow \gamma \geq \eta) \\
&= \|(1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}) - (\gamma - \eta)\nabla F(w_{k,t}^{\text{md},m})\|^2 \\
&\quad + (\gamma - \eta)^2 \mathbb{E}[\| \nabla F(w_{k,t}^{\text{md},m}) - g_{k,t}^m \|^2 | \mathcal{F}_{k,t}] \\
&\leq (1 - \alpha^{-1})^2 \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 + (\gamma - \eta)^2 \|\nabla F(w_{k,t}^{\text{md},m})\|^2 \\
&\quad + (\gamma - \eta)^2 \sigma^2 - 2(\gamma - \eta) \langle (1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}), \nabla F(w_{k,t}^{\text{md},m}) \rangle \\
&\leq (1 - \alpha^{-1})^2 (1 + \gamma\mu) \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 \\
&\quad + (\gamma - \eta)^2 (1 + \frac{1}{\gamma\mu}) \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + (\gamma - \eta)^2 \sigma^2 \\
&= \frac{(1 - \gamma\mu)^2}{1 + \gamma\mu} \|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + (\gamma - \eta)^2 \frac{1 + \gamma\mu}{\gamma\mu} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + (\gamma - \eta)^2 \sigma^2
\end{aligned}$$

Here, we need to bound  $\|\nabla F(w_{k,t}^{\text{md},m})\|^2$ .

$$\begin{aligned}
\|\nabla F(w_{k,t}^{\text{md},m})\|^2 &\leq 2L(F(w_{k,t}^{\text{md},m}) - F^*) \quad (\because \text{Assumption 3}) \\
&\leq 2L\left(\beta^{-1}(F(w_{k,t}^m) - F(w^*)) + (1 - \beta^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*)\right) \\
&\leq \beta^{-1}L^2\|w_{k,t}^m - w^*\|^2 + 2(1 - \beta^{-1})L(F(w_{k,t}^{\text{ag},m}) - F^*) \\
&= \frac{\gamma\mu L^2}{1 + \gamma\mu}\|w_{k,t}^m - w^*\|^2 + \frac{2L}{1 + \gamma\mu}(F(w_{k,t}^{\text{ag},m}) - F^*) \\
&\leq \frac{\mu L}{1 + \gamma\mu}\|w_{k,t}^m - w^*\|^2 + \frac{2L}{1 + \gamma\mu}(F(w_{k,t}^{\text{ag},m}) - F^*) = \frac{2L}{1 + \gamma\mu}\Psi_{k,t}^m \quad (38)
\end{aligned}$$

The last inequality comes from the fact  $\gamma \in [0, \frac{1}{L}]$ . Therefore, we finally get

$$\begin{aligned}
\mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] &\leq \frac{(1 - \gamma\mu)^2}{1 + \gamma\mu}\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + (\gamma - \eta)^2 \frac{1 + \gamma\mu}{\gamma\mu} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + (\gamma - \eta)^2 \sigma^2 \\
&\leq \frac{(1 - \gamma\mu)^2}{1 + \gamma\mu}\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + (\gamma - \eta)^2 \frac{1 + \gamma\mu}{\gamma\mu} \left( \frac{2L}{1 + \gamma\mu} \Psi_{k,t}^m \right) + (\gamma - \eta)^2 \sigma^2 \quad (39)
\end{aligned}$$

Now, let's compute  $\mathbb{E}[\Psi_{k,t+1}^m | \mathcal{F}_{k,t}]$ . We need to compute  $\mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}]$  and  $\mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}]$  first.

$$\begin{aligned}
\mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}] &= \mathbb{E}[\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma g_{k,t}^m - w^*\|^2 | \mathcal{F}_{k,t}] \\
&\leq \|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*\|^2 + \gamma^2 \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \gamma^2 \sigma^2 \\
&\quad - 2\gamma \langle (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle \\
&\leq (1 - \alpha^{-1})\|w_{k,t}^m - w^*\|^2 + \alpha^{-1}\|w_{k,t}^{\text{md},m} - w^*\|^2 + \gamma^2 \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \gamma^2 \sigma^2 \\
&\quad - 2\gamma \langle (1 - \alpha^{-1}(1 - \beta^{-1}))w_{k,t}^m + \alpha^{-1}(1 - \beta^{-1})w_{k,t}^{\text{ag},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle \\
&= (1 - \gamma\mu)\|w_{k,t}^m - w^*\|^2 + \gamma\mu\|w_{k,t}^{\text{md},m} - w^*\|^2 + \gamma^2 \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \gamma^2 \sigma^2 \\
&\quad - 2\gamma \left\langle \frac{1}{1 + \gamma\mu}w_{k,t}^m + \frac{\gamma\mu}{1 + \gamma\mu}w_{k,t}^{\text{ag},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \right\rangle
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}] &\leq \mathbb{E}[F(w_{k,t}^{\text{md},m}) + \langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t+1}^{\text{ag},m} - w_{k,t}^{\text{md},m} \rangle + \frac{L}{2}\|w_{k,t+1}^{\text{ag},m} - w_{k,t}^{\text{md},m}\|^2 - F^* | \mathcal{F}_{k,t}] \\
&\leq F(w_{k,t}^{\text{md},m}) - F^* - \eta \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2} \sigma^2 \\
&\leq F(w_{k,t}^{\text{md},m}) - F^* - \frac{\eta}{2} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2} \sigma^2 \quad (\because 1 - \frac{\eta L}{2} \geq \frac{1}{2} \leftarrow \eta \in [0, \frac{1}{L}]) \\
&= (1 - \alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) + \alpha^{-1}(F(w_{k,t}^{\text{md},m}) - F^*) \\
&\quad + (1 - \alpha^{-1})(F(w_{k,t}^{\text{md},m}) - F(w_{k,t}^{\text{ag},m})) - \frac{\eta}{2} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2} \sigma^2
\end{aligned}$$

$$\begin{aligned}
&\leq (1 - \alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) - \frac{\mu\alpha^{-1}}{2}\|w_{k,t}^{\text{md},m} - w^*\|^2 + \alpha^{-1}\langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t}^{\text{md},m} - w^* \rangle \\
&+ (1 - \alpha^{-1})\langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t}^{\text{md},m} - w_{k,t}^{\text{ag},m} \rangle - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&= (1 - \alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) - \frac{\mu\alpha^{-1}}{2}\|w_{k,t}^{\text{md},m} - w^*\|^2 - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&+ \alpha^{-1}\langle \nabla F(w_{k,t}^{\text{md},m}), \alpha\beta^{-1}w_{k,t}^m + (1 - \alpha\beta^{-1})w_{k,t}^{\text{ag},m} - w^* \rangle \\
&= (1 - \gamma\mu)(F(w_{k,t}^{\text{ag},m}) - F^*) - \frac{\gamma\mu^2}{2}\|w_{k,t}^{\text{md},m} - w^*\|^2 - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&+ \gamma\mu\left\langle \frac{1}{1 + \gamma\mu}w_{k,t}^m + \frac{\gamma\mu}{1 + \gamma\mu}w_{k,t}^{\text{ag},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \right\rangle
\end{aligned}$$

Then, we bound  $\mathbb{E}[\Psi_{k,t+1}^m | \mathcal{F}_{k,t}]$  by using the above results.

$$\begin{aligned}
\mathbb{E}[\Psi_{k,t+1}^m | \mathcal{F}_{k,t}] &= \frac{\mu}{2}\mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}] + \mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}] \\
&\leq (1 - \gamma\mu)\Psi_{k,t}^m - \frac{\eta - \gamma^2\mu}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\gamma^2\mu + \eta^2 L}{2}\sigma^2 \\
&\leq (1 - \gamma\mu)\Psi_{k,t}^m + \frac{\gamma^2\mu + \eta^2 L}{2}\sigma^2 \quad (\because \gamma \leq \sqrt{\frac{\eta}{\mu}}) \\
&\leq (1 - \gamma\mu)\Psi_{k,t}^m + \frac{\gamma^2(\mu + L)}{2}\sigma^2
\end{aligned} \tag{40}$$

Plugging (39), (40) in (37) yields,

$$\begin{aligned}
&\mathbb{E}[A_{k,t+1}^m | \mathcal{F}_{k,t}] \\
&\leq \frac{\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} \left( \frac{(1 - \gamma\mu)^2}{1 + \gamma\mu}\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + (\gamma - \eta)^2 \frac{1 + \gamma\mu}{\gamma\mu} \left( \frac{2L}{1 + \gamma\mu}\Psi_{k,t}^m \right) + (\gamma - \eta)^2\sigma^2 \right) \\
&+ \gamma^2(\mu + L) \frac{2L}{1 + \gamma\mu} \left( (1 - \gamma\mu)\Psi_{k,t}^m + \frac{\gamma^2(\mu + L)}{2}\sigma^2 \right) \\
&= \frac{(1 - \gamma\mu)^2}{1 + \gamma\mu} \cdot \frac{\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2}\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + \left( \frac{\gamma\mu(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \gamma^2(\mu + L)(1 - \gamma\mu) \right) \frac{2L}{1 + \gamma\mu}\Psi_{k,t}^m \\
&+ \left( \frac{\gamma^2\mu^2(\gamma - \eta)^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2 L}{1 + \gamma\mu} \right) \sigma^2
\end{aligned} \tag{41}$$

Since  $\eta \leq \gamma$ , we get  $(\gamma - \eta)^2 \leq \gamma^2$ . By using this fact, we obtain

$$\begin{aligned}
\frac{\gamma\mu(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \gamma^2(\mu + L)(1 - \gamma\mu) &\leq \frac{\gamma^3\mu(\mu + L)}{1 + \gamma\mu} + \gamma^2(\mu + L)(1 - \gamma\mu) \\
&= \gamma^2(\mu + L) \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)
\end{aligned} \tag{42}$$

It is easy to show that  $1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} < 1$ . Also, we get

$$\frac{(1 - \gamma\mu)^2}{1 + \gamma\mu} < 1 - \gamma\mu < 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \tag{43}$$

From (41), (42), and (43) we finally get

$$\mathbb{E}[A_{k,t+1}^m | \mathcal{F}_{k,t}] \leq (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})A_{k,t}^m + \left( \frac{\gamma^2\mu^2(\gamma - \eta)^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2L}{1 + \gamma\mu} \right)\sigma^2$$

From this relationship between  $A_{k,t+1}^m$  and  $A_{k,t}^m$ , we obtain the result of Proposition 14.

$$\begin{aligned} & \mathbb{E}[A_{k,t}^m] \\ & \leq (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t \mathbb{E}[A_{k,0}^m] + \left( \frac{\gamma^2\mu^2(\gamma - \eta)^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2L}{1 + \gamma\mu} \right)\sigma^2 \cdot \frac{1 - (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t}{1 - (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})} \\ & = (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t \mathbb{E}[A_{k,0}^m] + \left( \frac{(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \frac{\gamma^2(\mu + L)^2L}{\mu^2} \right)\sigma^2 \cdot \left( 1 - (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t \right) \\ & \leq \mathbb{E}[A_{k,0}^m] + \left( \frac{(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \frac{\gamma^2(\mu + L)^2L}{\mu^2} \right) \cdot \left( 1 - (1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu})^t \right)\sigma^2 \end{aligned}$$

**Proposition 15:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{1}{\gamma\mu}, \beta = \alpha + 1, \gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}], \eta, \gamma \in (0, \frac{1}{L}], \tau \geq 2$ , FedAQ yields

$$\begin{aligned} \frac{\mu}{2}\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2}\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] & \leq \left( \frac{4\gamma^2\mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right)\tau^2\mathbb{E}[\|\Psi_k\|] + (\gamma^2\mu + \eta^2L)\tau\sigma^2 \\ & \quad + \left( \frac{(\gamma - \eta)^2\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2L}{1 + \gamma\mu} \right)\frac{\tau^3\sigma^2}{2} \end{aligned}$$

Proof of Proposition 15 Let's first bound  $\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2]$  and  $\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2]$  individually.

$$\begin{aligned} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] & = \mathbb{E}[\|(w_{k,\tau}^m - w_{k,\tau-1}^m) + \dots + (w_{k,1}^m - w_{k,0}^m)\|^2] \\ & = \mathbb{E}\left[\left\|\sum_{t=0}^{\tau-1} \left( (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w_{k,t}^m - \gamma g_{k,t}^m \right)\right\|^2\right] \\ & = \mathbb{E}\left[\left\|\alpha^{-1}\sum_{t=0}^{\tau-1} (w_{k,t}^{\text{md},m} - w_{k,t}^m) - \gamma\sum_{t=0}^{\tau-1} g_{k,t}^m\right\|^2\right] \\ & \leq 2\alpha^{-2}\mathbb{E}\left[\left\|\sum_{t=0}^{\tau-1} (w_{k,t}^{\text{md},m} - w_{k,t}^m)\right\|^2\right] + 2\gamma^2\mathbb{E}\left[\left\|\sum_{t=0}^{\tau-1} g_{k,t}^m\right\|^2\right] \\ & \leq 2\alpha^{-2}\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|w_{k,t}^{\text{md},m} - w_{k,t}^m\|^2] + 2\gamma^2\mathbb{E}\left[\left\|\sum_{t=0}^{\tau-1} \nabla F(w_{k,t}^{\text{md},m})\right\|^2\right] \\ & \quad + 2\gamma^2\mathbb{E}\left[\left\|\sum_{t=0}^{\tau-1} (g_{k,t}^m - \nabla F(w_{k,t}^{\text{md},m}))\right\|^2\right] \\ & \leq 2\alpha^{-2}(1 - \beta^{-1})^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\gamma^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2] \\ & \quad + 2\gamma^2\sum_{t=0}^{\tau-1}\mathbb{E}[\|g_{k,t}^m - \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ & = \tau\left(\sum_{t=0}^{\tau-1} 2\alpha^{-2}(1 - \beta^{-1})^2\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\gamma^2\mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2]\right) + 2\tau\gamma^2\sigma^2 \end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] &= \mathbb{E}[\|\sum_{t=0}^{\tau-1} (w_{k,t+1}^{\text{ag},m} - w_{k,t}^{\text{ag},m})\|^2] \\
&= \mathbb{E}[\|\sum_{t=0}^{\tau-1} (w_{k,t}^{\text{md},m} - w_{k,t}^{\text{ag},m} - \eta g_{k,t}^m)\|^2] \\
&\leq 2\mathbb{E}[\|\sum_{t=0}^{\tau-1} (w_{k,t}^{\text{md},m} - w_{k,t}^{\text{ag},m})\|^2] + 2\eta^2\mathbb{E}[\|\sum_{t=0}^{\tau-1} g_{k,t}^m\|^2] \\
&= 2\beta^{-2}\mathbb{E}[\|\sum_{t=0}^{\tau-1} (w_{k,t}^m - w_{k,t}^{\text{ag},m})\|^2] + 2\eta^2\mathbb{E}[\|\sum_{t=0}^{\tau-1} \nabla F(w_{k,t}^{\text{md},m})\|^2] \\
&\quad + 2\eta^2\mathbb{E}[\|\sum_{t=0}^{\tau-1} (g_{k,t}^m - \nabla F(w_{k,t}^{\text{md},m}))\|^2] \\
&\leq 2\beta^{-2}\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\eta^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2] \\
&\quad + 2\eta^2\sum_{t=0}^{\tau-1}\mathbb{E}[\|g_{k,t}^m - \nabla F(w_{k,t}^{\text{md},m})\|^2] \\
&= \tau\left(\sum_{t=0}^{\tau-1} 2\beta^{-2}\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\eta^2\mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2]\right) + 2\tau\eta^2\sigma^2
\end{aligned}$$

Thus, by using the above results, we get

$$\begin{aligned}
&\frac{\mu}{2}\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2}\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
&\leq \tau\sum_{t=0}^{\tau-1}\left\{\left(\mu\alpha^{-2}(1-\beta^{-1})^2 + L\beta^{-2}\right)\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + (\gamma^2\mu + \eta^2L)\mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2]\right\} \\
&\quad + (\gamma^2\mu + \eta^2L)\tau\sigma^2 \\
&\leq \tau\sum_{t=0}^{\tau-1}\left\{\left(\mu\alpha^{-2}(1-\beta^{-1})^2 + L\beta^{-2}\right)\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + (\gamma^2\mu + \eta^2L)\frac{2L}{1+\gamma\mu}\mathbb{E}[\Psi_{k,t}^m]\right\} \\
&\quad + (\gamma^2\mu + \eta^2L)\tau\sigma^2 (\because (38)) \\
&\leq \tau\sum_{t=0}^{\tau-1}\left\{\frac{\gamma^2\mu^2(\mu+L)}{(1+\gamma\mu)^2}\mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + \gamma^2(\mu+L)\frac{2L}{1+\gamma\mu}\mathbb{E}[\Psi_{k,t}^m]\right\} + (\gamma^2\mu + \eta^2L)\tau\sigma^2 \\
&= \tau\left(\sum_{t=0}^{\tau-1}\mathbb{E}[A_{k,t}^m]\right) + (\gamma^2\mu + \eta^2L)\tau\sigma^2
\end{aligned}$$



By Proposition 14 and the fact  $\Psi_{k,0}^m = \Psi_k$ , we obtain

$$\begin{aligned}
& \frac{\mu}{2} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2} \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
& \leq \tau \left\{ \sum_{t=0}^{\tau-1} \mathbb{E}[A_{k,0}^m] + \left( \frac{(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \frac{\gamma^2(\mu + L)^2 L}{\mu^2} \right) \cdot \left( 1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^t \right) \sigma^2 \right\} \\
& \quad + (\gamma^2\mu + \eta^2 L) \tau \sigma^2 \\
& = \tau^2 \left( \frac{\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} \mathbb{E}[\|w_k - w_k^{\text{ag}}\|^2] + \gamma^2(\mu + L) \frac{2L}{1 + \gamma\mu} \mathbb{E}[\Psi_k] \right) \\
& \quad + \tau \left( \frac{(\gamma - \eta)^2(\mu + L)}{1 + \gamma\mu} + \frac{\gamma^2(\mu + L)^2 L}{\mu^2} \right) \left( \sum_{t=0}^{\tau-1} 1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^t \right) \sigma^2 + (\gamma^2\mu + \eta^2 L) \tau \sigma^2
\end{aligned}$$

Before we get to the final result, let's find the upper bound for  $\|w_k - w_k^{\text{ag}}\|^2$ ,  $\sum_{t=0}^{\tau-1} \left( 1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^t \right)$

$$\begin{aligned}
\|w_k - w_k^{\text{ag}}\|^2 &= \|w_k - w^* - (w_k^{\text{ag}} - w^*)\|^2 \\
&\leq 2\|w_k - w^*\|^2 + 2\|w_k^{\text{ag}} - w^*\|^2 \\
&\leq 2\|w_k - w^*\|^2 + 2 \cdot \frac{2}{\mu} \left( F(w_k^{\text{ag}}) - F^* - \langle \nabla F(w^*), w_k^{\text{ag}} - w^* \rangle \right) \\
&= 2\|w_k - w^*\|^2 + \frac{4}{\mu} (F(w_k^{\text{ag}}) - F^*) = \frac{4}{\mu} \Psi_k
\end{aligned}$$

$$\begin{aligned}
\sum_{t=0}^{\tau-1} \left( 1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^t \right) &= \tau - \sum_{t=0}^{\tau-1} \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^t \\
&= \tau - \frac{1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)^\tau}{1 - \left( 1 - \gamma\mu + \frac{\gamma\mu}{1 + \gamma\mu} \right)} \\
&\leq \tau - \frac{1 - \left( 1 - \frac{\gamma^2\mu^2}{1 + \gamma\mu} \tau + \left( \frac{\gamma^2\mu^2}{1 + \gamma\mu} \right)^2 \frac{\tau(\tau-1)}{2} \right)}{\frac{\gamma^2\mu^2}{1 + \gamma\mu}} \\
&= \frac{\gamma^2\mu^2}{1 + \gamma\mu} \cdot \frac{\tau(\tau-1)}{2} \leq \frac{\gamma^2\mu^2}{1 + \gamma\mu} \cdot \frac{\tau^2}{2}
\end{aligned}$$

Therefore, we conclude as below

$$\begin{aligned}
\frac{\mu}{2} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2} \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] &\leq \left( \frac{4\gamma^2\mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right) \tau^2 \mathbb{E}[\Psi_k] + (\gamma^2\mu + \eta^2 L) \tau \sigma^2 \\
&\quad + \left( \frac{(\gamma - \eta)^2\gamma^2\mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2 L}{1 + \gamma\mu} \right) \frac{\tau^3 \sigma^2}{2}
\end{aligned}$$

**Proof of Lemma 10** By the definition of  $\Psi_k$ ,  $\Psi_{k,t}$  and Proposition 11,

$$\mathbb{E}[\Psi_{k+1}] = \mathbb{E}[\Psi_{k,\tau}] + \frac{\mu}{2} \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] + \mathbb{E}[F(w_{k+1}^{\text{ag}})] - \frac{1}{M} \sum_{m=1}^M F(w_{k,\tau}^{\text{ag},m})$$

Applying Proposition 12 and Proposition 13, we have

$$\begin{aligned}
& \mathbb{E}[\Psi_{k+1}] \\
& \leq (1 - \gamma\mu)^\tau \mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2 L + \frac{\gamma^2 \mu}{M})\tau\sigma^2 \\
& + \gamma\mu L\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \|\frac{1}{1 + \gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})\|] \\
& + \frac{q\mu}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{qL}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
& \leq (1 - \gamma\mu)^\tau \mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2 L + \frac{\gamma^2 \mu}{M})\tau\sigma^2 \\
& + \gamma\mu L\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \|\frac{1}{1 + \gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})\|] \\
& + \frac{q}{M} \left[ \left( \frac{4\gamma^2 \mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right) \tau^2 \mathbb{E}[\Psi_k] + (\gamma^2 \mu + \eta^2 L)\tau\sigma^2 \right. \\
& \left. + \left( \frac{(\gamma - \eta)^2 \gamma^2 \mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2 L}{1 + \gamma\mu} \right) \frac{\tau^3 \sigma^2}{2} \right] \\
& = \left\{ (1 - \gamma\mu)^\tau + \frac{q}{M} \left( \frac{4\gamma^2 \mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right) \tau^2 \right\} \mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2 L + \frac{\gamma^2 \mu}{M})\tau\sigma^2 \\
& + \frac{q}{M}(\gamma^2 \mu + \eta^2 L)\tau\sigma^2 + \frac{q}{2M} \left( \frac{(\gamma - \eta)^2 \gamma^2 \mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2 L}{1 + \gamma\mu} \right) \tau^3 \sigma^2 \\
& + \gamma\mu L\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \|\frac{1}{1 + \gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})\|]
\end{aligned}$$

The second inequality comes from Proposition 15. Then, let's define  $C(\gamma, \tau)$  as

$$C(\gamma, \tau) = (1 - \gamma\mu)^\tau + \frac{q}{M} \left( \frac{4\gamma^2 \mu(\mu + L)}{(1 + \gamma\mu)^2} + \frac{2L\gamma^2(\mu + L)}{1 + \gamma\mu} \right) \tau^2$$

Finally, we obtain

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] & \leq C(\gamma, \tau) \mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2 L + \frac{\gamma^2 \mu}{M})\tau\sigma^2 + \frac{q}{M}(\gamma^2 \mu + \eta^2 L)\tau\sigma^2 \\
& + \frac{q}{2M} \left( \frac{(\gamma - \eta)^2 \gamma^2 \mu^2(\mu + L)}{(1 + \gamma\mu)^2} + \frac{\gamma^4(\mu + L)^2 L}{1 + \gamma\mu} \right) \tau^3 \sigma^2 + \gamma\mu L\tau \\
& \cdot \max_{0 \leq t < \tau} \mathbb{E}[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \|\frac{1}{1 + \gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1 + \gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m})\|]
\end{aligned}$$

### 5.3.2 Proof of Theorem 16

**Theorem 16:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 1, 2, 3, 4, then for  $\alpha = \frac{1}{\gamma\mu}, \beta = \alpha + 1, \gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}}), \eta, \gamma \in (0, \frac{1}{L}), \tau \geq 2$ , if the learning rate  $\gamma$  satisfies

$$\left( \mu^2 + \frac{q}{M}(\mu + L)(4\mu + 2L) \right) \gamma\tau \leq \frac{1}{2}\mu \quad (44)$$

FedAQ yields

$$\begin{aligned} \mathbb{E}[\Psi_K] &\leq \exp\left(-\frac{1}{2}\max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}})K\tau\right)\Psi_0 + (2q+1)\left(\frac{\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}M\tau^{\frac{1}{2}}} + \frac{\eta\sigma^2}{M}\right) + 14\eta^2L\tau\sigma^2 \\ &\quad + \frac{(780 + \frac{2q}{M})\eta^{\frac{3}{2}}L\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}} + \frac{(\mu+L)(\mu^2 + \mu L + L^2)q\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{5}{2}}M} + \frac{q\eta^3\tau^2(\mu+L)^2L\sigma^2}{\mu M} \end{aligned}$$

Proof of Theorem 16 At first, due to the condition (44) in Theorem 16, we get

$$\begin{aligned} C(\gamma, \tau) &= (1 - \gamma\mu)^\tau + \frac{q}{M}\left(\frac{4\gamma^2\mu(\mu+L)}{(1+\gamma\mu)^2} + \frac{2L\gamma^2(\mu+L)}{1+\gamma\mu}\right)\tau^2 \\ &\leq 1 - \gamma\mu\tau + \gamma^2\mu^2\tau^2 + \frac{q}{M}\gamma^2(\mu+L)(4\mu+2L)\tau^2 \\ &= 1 - \gamma\mu\tau + \left(\mu^2 + \frac{q}{M}(\mu+L)(4\mu+2L)\right)\gamma^2\tau^2 \\ &\leq 1 - \frac{1}{2}\gamma\mu\tau \quad (\because \text{condition (44)}) \end{aligned}$$

The first inequality comes from the fact that  $(1 - \gamma\mu)^\tau \leq e^{-\gamma\mu\tau} \leq 1 - \gamma\mu\tau + \gamma^2\mu^2\tau^2$  when  $0 \leq \gamma\mu \leq 1$ . Also, it is trivial that  $\gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}}) \in [\eta, \sqrt{\frac{\eta}{\mu\tau}}]$ . Thus, we can use Lemma 10. By using Lemma 10 and the above result, we obtain

$$\begin{aligned} \mathbb{E}[\Psi_{k+1}] &\leq (1 - \frac{1}{2}\gamma\mu\tau)\mathbb{E}[\Psi_k] + \frac{1}{2}(\eta^2L + \frac{\gamma^2\mu}{M})\tau\sigma^2 \\ &\quad + \frac{q}{M}(\gamma^2\mu + \eta^2L)\tau\sigma^2 + \frac{q}{2M}\left(\frac{(\gamma-\eta)^2\gamma^2\mu^2(\mu+L)}{(1+\gamma\mu)^2} + \frac{\gamma^4(\mu+L)^2L}{1+\gamma\mu}\right)\tau^3\sigma^2 + \gamma\mu L\tau \\ &\quad \cdot \max_{0 \leq t < \tau} \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \left\| \frac{1}{1+\gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1+\gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m}) \right\| \right] \quad (45) \end{aligned}$$

By the Lemma B.3 in [43], we know that the below quantity is bounded.

$$\begin{aligned} \max_{0 \leq t < \tau} \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^M \|\bar{w}_{k,t}^{\text{md}} - w_{k,t}^{\text{md},m}\| \left\| \frac{1}{1+\gamma\mu}(\bar{w}_{k,t} - w_{k,t}^m) + \frac{\gamma\mu}{1+\gamma\mu}(\bar{w}_{k,t}^{\text{ag}} - w_{k,t}^{\text{ag},m}) \right\| \right] &\leq B \\ B &= \begin{cases} 7\eta\gamma\tau\sigma^2 \left(1 + \frac{2\gamma^2\mu}{\eta}\right)^{2\tau}, & \text{if } \gamma \in \left(\eta, \sqrt{\frac{\eta}{\mu}}\right] \\ 7\eta^2\tau\sigma^2, & \text{if } \gamma = \eta \end{cases} \end{aligned}$$

Telescoping (45) yields

$$\begin{aligned} \mathbb{E}[\Psi_K] &\leq (1 - \frac{1}{2}\gamma\mu\tau)^K\Psi_0 + \left(\sum_{k'=0}^{K-1} (1 - \frac{1}{2}\gamma\mu\tau)^{k'}\right) \cdot \left[\frac{1}{2}(\eta^2L + \frac{\gamma^2\mu}{M})\tau\sigma^2 + \gamma\mu L\tau B \right. \\ &\quad \left. + \frac{q}{M}(\gamma^2\mu + \eta^2L)\tau\sigma^2 + \frac{q}{2M}\left(\frac{(\gamma-\eta)^2\gamma^2\mu^2(\mu+L)}{(1+\gamma\mu)^2} + \frac{\gamma^4(\mu+L)^2L}{1+\gamma\mu}\right)\tau^3\sigma^2\right] \\ &\leq \exp\left(-\frac{\gamma\mu\tau K}{2}\right)\Psi_0 + \frac{\eta^2L\sigma^2}{\gamma\mu} + \frac{\gamma\sigma^2}{M} + 2LB + 2q\left(\frac{\gamma\sigma^2}{M} + \frac{\eta^2L\sigma^2}{\gamma\mu M}\right) \\ &\quad + \frac{q}{M}\left(\frac{(\gamma-\eta)^2\gamma\mu(\mu+L)}{(1+\gamma\mu)^2} + \frac{\gamma^3(\mu+L)^2L}{(1+\gamma\mu)\mu}\right)\tau^2\sigma^2 \end{aligned}$$

The last inequality comes from the fact that  $\sum_{k'=0}^{K-1} (1 - \frac{1}{2}\gamma\mu\tau)^{k'} \leq \frac{2}{\gamma\mu\tau}$ . Since we plug in  $\gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}})$ , we can use Lemma B.4 in [43]. Therefore, we obtain

$$\begin{aligned} \mathbb{E}[\Psi_K] &\leq \exp\left(-\frac{1}{2}\max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}})K\tau\right)\Psi_0 + \frac{\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}M\tau^{\frac{1}{2}}} + \frac{\eta\sigma^2}{M} + \frac{780\eta^{\frac{3}{2}}L\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}} + 14\eta^2L\tau\sigma^2 \\ &+ \max\left(\frac{2q\eta^{\frac{1}{2}}\sigma^2}{M\mu^{\frac{1}{2}}\tau^{\frac{1}{2}}}, \frac{2q\eta\sigma^2}{M}\right) + \min\left(\frac{2q\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}L\sigma^2}{M\mu^{\frac{1}{2}}}, \frac{2q\eta L\sigma^2}{M\mu}\right) \\ &+ \frac{q\tau^2\sigma^2}{M} \max\left(\frac{\eta^{\frac{3}{2}}\mu(\mu+L)}{\mu^{\frac{3}{2}}\tau^{\frac{3}{2}}}, \frac{\eta^{\frac{3}{2}}(\mu+L)^2L}{\mu^{\frac{5}{2}}\tau^{\frac{3}{2}}}, \frac{\eta^3(\mu+L)^2L}{\mu}\right) \end{aligned}$$

The first term stems directly from Lemma B.4 in [43]. Also, the last term comes from the fact that

$$\frac{(\gamma - \eta)^2\gamma\mu(\mu+L)}{(1+\gamma\mu)^2} + \frac{\gamma^3(\mu+L)^2L}{(1+\gamma\mu)\mu} \leq \begin{cases} \gamma^3\mu(\mu+L) + \frac{\gamma^3(\mu+L)^2L}{\mu}, & \text{if } \gamma \neq \eta \\ \frac{\eta^3(\mu+L)^2L}{\mu}, & \text{if } \gamma = \eta \end{cases}$$

Therefore, by simple inequalities such as  $\max(a, b) \leq a + b$  and  $\min(a, b) \leq a$ , we ultimately get

$$\begin{aligned} \mathbb{E}[\Psi_K] &\leq \exp\left(-\frac{1}{2}\max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}})K\tau\right)\Psi_0 + \frac{(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}M\tau^{\frac{1}{2}}} + \frac{(2q+1)\eta\sigma^2}{M} + 14\eta^2L\tau\sigma^2 \\ &+ \frac{(780 + \frac{2q}{M})\eta^{\frac{3}{2}}L\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}} + \frac{(\mu+L)(\mu^2 + \mu L + L^2)q\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{5}{2}}M} + \frac{q\eta^3\tau^2(\mu+L)^2L\sigma^2}{\mu M} \end{aligned} \quad (46)$$

### 5.3.3 Proof of Corollary 17

**Corollary 17:** Let  $C_1, C_2$ , and  $\eta_0$  as below. Note that  $T = K\tau$ .

$$\begin{aligned} C_1 &= \frac{(\mu+L)(\mu^2 + \mu L + L^2)q}{\mu^{\frac{5}{2}}}, \quad C_2 = \frac{q(\mu+L)^2L}{\mu} \\ \eta_0 &= \frac{4\tau}{\mu T^2} \log^2\left(e + \min\left(\frac{\mu MT\Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right)\right) \end{aligned}$$

Then for  $\eta = \min(\frac{1}{L}, \eta_0)$ , FedAQ yields

$$\begin{aligned} \mathbb{E}[\Psi_K] &\leq \min\left(\exp\left(-\frac{\mu T}{2L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}}T}{2L^{\frac{1}{2}}\tau^{\frac{1}{2}}}\right)\right)\Psi_0 \\ &+ \frac{7(2q+1)\sigma^2}{\mu MT} \log^2\left(e + \frac{\mu MT\Psi_0}{(2q+1)\sigma^2}\right) \end{aligned} \quad (47)$$

$$+ \frac{(6465 + \frac{16q}{M})L\tau^2\sigma^2}{\mu^2 T^3} \log^4\left(e + \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}\right) \quad (48)$$

$$+ \frac{9(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}{\mu^3 MT^3} \log^6\left(e + \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right) \quad (49)$$

Proof of Corollary 17 Let's decompose the final result (46) of the Theorem 16 into a decreasing term and an

increasing term. We denote the decreasing term  $\psi_1$  and the increasing term  $\psi_2$  as below.

$$\begin{aligned}\psi_1(\eta) &= \exp\left(-\frac{1}{2}\max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}}T)\Psi_0\right) \\ \psi_2(\eta) &= \frac{(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}M\tau^{\frac{1}{2}}} + \frac{(2q+1)\eta\sigma^2}{M} + \frac{(780 + \frac{2q}{M})\eta^{\frac{3}{2}}L\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}}} + 14\eta^2L\tau\sigma^2 \\ &\quad + \frac{(\mu+L)(\mu^2 + \mu L + L^2)q\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{5}{2}}M} + \frac{q\eta^3\tau^2(\mu+L)^2L\sigma^2}{\mu M}\end{aligned}$$

Since  $\psi_1$  is the decreasing term, we have

$$\psi_1(\eta) \leq \psi_1\left(\frac{1}{L}\right) + \psi_1(\eta_0) \quad (50)$$

where

$$\begin{aligned}\psi_1\left(\frac{1}{L}\right) &= \min\left(\exp\left(-\frac{\mu T}{2L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}}T}{2L^{\frac{1}{2}}\tau^{\frac{1}{2}}}\right)\right)\Psi_0 \\ \psi_1(\eta_0) &\leq \exp\left(-\frac{1}{2}\sqrt{\frac{\eta_0\mu}{\tau}}T\right) \\ &= \left(e + \min\left(\frac{\mu MT\Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right)\right)^{-1}\Psi_0 \\ &\leq \frac{(2q+1)\sigma^2}{\mu MT} + \frac{L\tau^2\sigma^2}{\mu^2 T^3} + \frac{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}{\mu^3 MT^3}\end{aligned}$$

Since  $\psi_2$  is the increasing term, we have

$$\begin{aligned}\psi_2(\eta) &\leq \psi_2(\eta_0) \\ &\leq \frac{2(2q+1)\sigma^2}{\mu MT} \log\left(e + \frac{\mu MT\Psi_0}{(2q+1)\sigma^2}\right) + \frac{4(2q+1)\tau\sigma^2}{\mu MT^2} \log^2\left(e + \frac{\mu MT\Psi_0}{(2q+1)\sigma^2}\right) \\ &\quad + \frac{8(780 + \frac{2q}{M})L\tau^2\sigma^2}{\mu^2 T^3} \log^3\left(e + \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}\right) + \frac{224L\tau^3\sigma^2}{\mu^2 T^4} \log^4\left(e + \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}\right) \\ &\quad + \frac{8C_1\tau^2\sigma^2}{\mu^{\frac{3}{2}}MT^3} \log^3\left(e + \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right) + \frac{64C_2\tau^5\sigma^2}{\mu^3 MT^6} \log^6\left(e + \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right) \\ &\leq \frac{6(2q+1)\sigma^2}{\mu MT} \log^2\left(e + \frac{\mu MT\Psi_0}{(2q+1)\sigma^2}\right) + \frac{(6464 + \frac{16q}{M})L\tau^2\sigma^2}{\mu^2 T^3} \log^4\left(e + \frac{\mu^2 T^3\Psi_0}{L\tau^2\sigma^2}\right) \\ &\quad + \frac{8(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}{\mu^3 MT^3} \log^6\left(e + \frac{\mu^3 MT^3\Psi_0}{(\mu^{\frac{3}{2}}C_1 + 8C_2)\tau^2\sigma^2}\right)\end{aligned} \quad (51)$$

The last inequality comes from  $\frac{\tau}{T} \leq 1$ . Therefore, by combining (50) and (51), we finally get

$$\begin{aligned}
\mathbb{E}[\Psi_K] &\leq \psi_1(\eta) + \psi_2(\eta) \\
&\leq \psi_1\left(\frac{1}{L}\right) + \psi_1(\eta_0) + \psi_2(\eta_0) \\
&\leq \min\left(\exp\left(-\frac{\mu T}{2L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}} T}{2L^{\frac{1}{2}} \tau^{\frac{1}{2}}}\right)\right) \Psi_0 + \frac{7(2q+1)\sigma^2}{\mu MT} \log^2\left(e + \frac{\mu MT \Psi_0}{(2q+1)\sigma^2}\right) \\
&\quad + \frac{(6465 + \frac{16q}{M})L\tau^2\sigma^2}{\mu^2 T^3} \log^4\left(e + \frac{\mu^2 T^3 \Psi_0}{L\tau^2\sigma^2}\right) \\
&\quad + \frac{9(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}{\mu^3 MT^3} \log^6\left(e + \frac{\mu^3 MT^3 \Psi_0}{(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}\right)
\end{aligned}$$

### 5.3.4 Why the Condition (44) is Satisfied

The synchronization rounds  $K$  required for linear speedup in  $M$  for FedAQ is  $\tilde{O}\left(\left(\frac{M}{1+q}\right)^{\frac{1}{2}}\right)$  (See Remark 8). Since we derive this result from Theorem 16, we should show that  $K = \tilde{O}\left(\left(\frac{M}{1+q}\right)^{\frac{1}{2}}\right)$  satisfies the condition (44) in Theorem 16.

$$\left(\mu^2 + \frac{q}{M}(\mu + L)(4\mu + 2L)\right)\gamma\tau \leq \frac{1}{2}\mu$$

We rewrite the above condition as below.

$$\gamma\tau \leq \frac{\mu}{2\mu^2 + \frac{2q}{M}(\mu + L)(4\mu + 2L)} \quad (52)$$

We know  $\gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}})$  and  $\eta = \min(\frac{1}{L}, \eta_0)$ . Since  $\eta_0$  becomes smaller and smaller as  $T$  increases, we assume  $\eta = \eta_0$  here. Therefore, we get

$$\begin{aligned}
\gamma\tau &= \max(\eta_0\tau, \sqrt{\frac{\eta_0\tau}{\mu}}) \\
&= \max\left(\frac{4\tau^2}{\mu T^2} \log^2\left(e + \min\left(\frac{\mu MT \Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3 \Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3 \Psi_0}{(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}\right)\right), \right. \\
&\quad \left. \frac{2\tau}{\mu T} \log\left(e + \min\left(\frac{\mu MT \Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3 \Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3 \Psi_0}{(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}\right)\right)\right)
\end{aligned}$$

Note that  $K = \frac{T}{\tau} = \tilde{O}\left(\left(\frac{M}{1+q}\right)^{\frac{1}{2}}\right) = C\left(\frac{M}{1+q}\right)^{\frac{1}{2}} \log(T)$  because  $\tilde{O}$  contains hidden multiplicative polylog factors with respect to  $T$ . We can assume  $T$  is sufficiently large here. Then, we have

$$\begin{aligned}
\gamma\tau &= \max\left(\frac{4(1+q)}{\mu C^2 M \log^2(T)} \log^2\left(e + \min\left(\frac{\mu MT \Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3 \Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3 \Psi_0}{(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}\right)\right), \right. \\
&\quad \left. \frac{2(1+q)^{\frac{1}{2}}}{\mu C M^{\frac{1}{2}} \log(T)} \log\left(e + \min\left(\frac{\mu MT \Psi_0}{(2q+1)\sigma^2}, \frac{\mu^2 T^3 \Psi_0}{L\tau^2\sigma^2}, \frac{\mu^3 MT^3 \Psi_0}{(\mu^{\frac{3}{2}} C_1 + 8C_2)\tau^2\sigma^2}\right)\right)\right) \\
&\leq \max\left(\frac{4(1+q)}{\mu C^2 M \log^2(T)} \log^2\left(\frac{2\mu MT \Psi_0}{(2q+1)\sigma^2}\right), \frac{2(1+q)^{\frac{1}{2}}}{\mu C M^{\frac{1}{2}} \log(T)} \log\left(\frac{2\mu MT \Psi_0}{(2q+1)\sigma^2}\right)\right)
\end{aligned}$$

For an arbitrary constant  $k_1 > 0$ , it is easy to show that  $\lim_{T \rightarrow \infty} \frac{\log(k_1 T)}{\log(T)} = 1$ . Thus, we obtain

$$\begin{aligned} \gamma\tau &\leq \max\left(\frac{4(1+q)}{\mu C^2 M \log^2(T)} \log^2\left(\frac{2\mu M T \Psi_0}{(2q+1)\sigma^2}\right), \frac{2(1+q)^{\frac{1}{2}}}{\mu C M^{\frac{1}{2}} \log(T)} \log\left(\frac{2\mu M T \Psi_0}{(2q+1)\sigma^2}\right)\right) \\ &\simeq \max\left(\frac{4(1+q)}{\mu C^2 M}, \frac{2(1+q)^{\frac{1}{2}}}{\mu C M^{\frac{1}{2}}}\right) \\ &\leq \frac{\mu}{2\mu^2 + \frac{2q}{M}(\mu + L)(4\mu + 2L)} \end{aligned}$$

Finally, we conclude that there exists a constant  $C$  that meets the last inequality. Therefore,  $K = \tilde{O}\left(\left(\frac{M}{1+q}\right)^{\frac{1}{2}}\right)$  satisfies the condition (44).

## 5.4 Proof Details for FedAQ under Condition Set (35)

We use notations defined in section 5.3 here as well. We newly define  $\Phi_{k,t}^m, \bar{\Phi}_{k,t}, \Phi_k, B_{k,t}^m$  as below.

$$\begin{aligned} \Phi_{k,t}^m &= F(w_{k,t}^{\text{ag},m}) - F^* + \frac{1}{6}\mu \|w_{k,t}^m - w^*\|^2 \\ \bar{\Phi}_{k,t} &= F(\bar{w}_{k,t}^{\text{ag}}) - F^* + \frac{1}{6}\mu \|\bar{w}_{k,t} - w^*\|^2 \\ \Phi_k &:= \Phi_{k,0} = F(w_k^{\text{ag}}) - F^* + \frac{1}{6}\mu \|w_k - w^*\|^2 \\ B_{k,t}^m &= \left(\frac{\mu\alpha^{-2}}{3}(1-\beta^{-1})^2 + L\beta^{-2}\right) \|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2 + \gamma^2 \left(\frac{\mu}{3} + L\right) \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\Phi_{k,t}^m \end{aligned}$$

The flow of proof is similar to section 5.3. We need one more condition  $\gamma\mu \leq \frac{3}{4}$  to show the convergence of FedAQ under the parameter condition set (35).

### 5.4.1 Proof of Lemma 5

In order to prove Lemma 5, we first introduce five crucial Propositions for proving Lemma 5. Then, we prove Lemma 5 by using Propositions in the last part of this section.

**Proposition 18:** Let Assumption 1 hold and consider any  $k$  synchronization round. Then, we can decompose the expectation as follows:

$$\begin{aligned} \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] + \mathbb{E}[\|\bar{w}_{k,\tau} - w^*\|^2] \\ \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F^*] &= \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F(\bar{w}_{k,\tau}^{\text{ag}})] + \mathbb{E}[F(\bar{w}_{k,\tau}^{\text{ag}}) - F^*] \end{aligned}$$

Proof of Proposition 18 The second equality is trivial. The first equality is the same as one in Proposition 11.

**Proposition 19:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{3}{2\gamma\mu} - \frac{1}{2}, \beta = \frac{2\alpha^2 - 1}{\alpha - 1}, \gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}], \eta \in (0, \frac{1}{L}]$ , FedAQ yields

$$\begin{aligned} \mathbb{E}[\Phi_{k,\tau}] &\leq (1 - \frac{1}{3}\gamma\mu)^\tau \mathbb{E}[\Phi_k] + \left(\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6}\right) \frac{\tau \sigma^2}{M} \\ &\quad + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \end{aligned}$$

Proof of Proposition 19 We refer to the proof of Lemma C.2 in [43]. There is no quantization between  $\Phi_{k,\tau}$  and  $\bar{\Phi}_k$ . Thus, we can directly apply useful inequalities in the proof of Lemma C.2 in [43] to our proof. Then, we obtain

$$\mathbb{E}[\Phi_{k,t+1}|\mathcal{F}_{k,t}] \leq (1 - \frac{1}{3}\gamma\mu)\Phi_{k,t} + (\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6})\frac{\sigma^2}{M} + \gamma\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M}\sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2$$

From the above relationship between  $\Phi_{k,t+1}$  and  $\Phi_{k,t}$ , we get

$$\begin{aligned} \mathbb{E}[\Phi_{k,\tau}] &\leq (1 - \frac{1}{3}\gamma\mu)^\tau \mathbb{E}[\Phi_k] + \left(\sum_{t=0}^{\tau-1} (1 - \frac{1}{3}\gamma\mu)^t\right) \cdot (\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6})\frac{\sigma^2}{M} \\ &\quad + \gamma \sum_{t=0}^{\tau-1} \left\{ (1 - \frac{1}{3}\gamma\mu)^{\tau-t-1} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M}\sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \right\} \\ &\leq (1 - \frac{1}{3}\gamma\mu)^\tau \mathbb{E}[\Phi_k] + (\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6})\frac{\tau\sigma^2}{M} \\ &\quad + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M}\sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \end{aligned}$$

**Proposition 20:** Let Assumption 1 hold. Then, we have

$$\begin{aligned} \mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] &\leq \frac{q}{M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] \\ \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F(\bar{w}_{k,\tau}^{\text{ag}})] &\leq \frac{qL}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \end{aligned}$$

Proof of Proposition 20 The first inequality is the same as one in Proposition 13. The proof of the second inequality is similar to Proposition 13 as well.

$$\begin{aligned} \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F(\bar{w}_{k,\tau}^{\text{ag}})] &= \mathbb{E}[F(w_k^{\text{ag}} + \frac{1}{M}\sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})) - F(\frac{1}{M}\sum_{m=1}^M w_{k,\tau}^{\text{ag},m})] \\ &\leq \mathbb{E}\left[\langle \nabla F(\frac{1}{M}\sum_{m=1}^M w_{k,\tau}^{\text{ag},m}), \frac{1}{M}\sum_{m=1}^M (Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) \right. \\ &\quad \left. - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})) \rangle + \frac{L}{2}\|\frac{1}{M}\sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2\right] \\ &= \frac{L}{2}\mathbb{E}[\|\frac{1}{M}\sum_{m=1}^M Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2] \\ &= \frac{L}{2M^2}\sum_{m=1}^M \mathbb{E}[\|Q(w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}) - (w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}})\|^2] \\ &\leq \frac{qL}{2M^2}\sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \end{aligned}$$



**Proposition 21:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{3}{2\gamma\mu} - \frac{1}{2}$ ,  $\beta = \frac{2\alpha^2-1}{\alpha-1}$ ,  $\gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}]$ ,  $\eta, \gamma \in (0, \frac{1}{L}]$ ,  $\gamma\mu \leq \frac{3}{4}$ , we get

$$\begin{aligned} \mathbb{E}[B_{k,t}^m] &\leq \mathbb{E}[B_{k,0}^m] + \left( \left( \frac{\mu}{3} \left( \frac{2\alpha-1}{2\alpha^2-1} \right)^2 + L \left( \frac{\alpha-1}{2\alpha^2-1} \right)^2 \right) \cdot (\gamma-\eta)^2 + \gamma^4 \left( \frac{\mu}{3} + L \right)^2 \frac{2\alpha^2-\alpha}{2\alpha^2-1} L \right) \\ &\quad \cdot \frac{1 + \frac{1}{2}\alpha^{-1}}{\frac{1}{4}\alpha^{-2}} \cdot \left( 1 - \left( 1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}} \right)^t \right) \sigma^2 \end{aligned}$$

Proof of Proposition 21 From the notation mentioned in the beginning of section 5.4,

$$\begin{aligned} \mathbb{E}[B_{k,t+1}^m | \mathcal{F}_{k,t}] &= \left( \frac{\mu\alpha^{-2}}{3} (1 - \beta^{-1})^2 + L\beta^{-2} \right) \mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] \\ &\quad + \gamma^2 \left( \frac{\mu}{3} + L \right) \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L \mathbb{E}[\Phi_{k,t+1}^m | \mathcal{F}_{k,t}] \end{aligned} \quad (53)$$

Thus, let's sequentially compute  $\mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}]$  and  $\mathbb{E}[\Phi_{k,t+1}^m | \mathcal{F}_{k,t}]$ .

$$\begin{aligned} \mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] &= \mathbb{E}[\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma g_{k,t}^m - w_{k,t}^{\text{md},m} + \eta g_{k,t}^m\|^2 | \mathcal{F}_{k,t}] \\ &= \mathbb{E}[\|(1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}) - (\gamma - \eta)g_{k,t}^m\|^2 | \mathcal{F}_{k,t}] \quad (\leftarrow \gamma \geq \eta) \\ &= \|(1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}) - (\gamma - \eta)\nabla F(w_{k,t}^{\text{md},m})\|^2 \\ &\quad + (\gamma - \eta)^2 \mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m}) - g_{k,t}^m\|^2 | \mathcal{F}_{k,t}] \\ &\leq (1 - \alpha^{-1})^2 \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 + (\gamma - \eta)^2 \|\nabla F(w_{k,t}^{\text{md},m})\|^2 \\ &\quad + (\gamma - \eta)^2 \sigma^2 - 2(\gamma - \eta) \langle (1 - \alpha^{-1})(w_{k,t}^m - w_{k,t}^{\text{md},m}), \nabla F(w_{k,t}^{\text{md},m}) \rangle \\ &\leq (1 - \alpha^{-1})^2 (1 + 2\alpha^{-1}) \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 \\ &\quad + (\gamma - \eta)^2 \left( 1 + \frac{\alpha}{2} \right) \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + (\gamma - \eta)^2 \sigma^2 \end{aligned}$$

Here, we need to bound  $\|\nabla F(w_{k,t}^{\text{md},m})\|^2$ .

$$\begin{aligned} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 &\leq 2L(F(w_{k,t}^{\text{md},m}) - F^*) \quad (\because \text{Assumption 3}) \\ &\leq 2L \left( \beta^{-1}(F(w_{k,t}^m) - F(w^*)) + (1 - \beta^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) \right) \\ &\leq \beta^{-1}L^2 \|w_{k,t}^m - w^*\|^2 + 2(1 - \beta^{-1})L(F(w_{k,t}^{\text{ag},m}) - F^*) \\ &= \frac{\alpha-1}{2\alpha^2-1} L^2 \|w_{k,t}^m - w^*\|^2 + 2L \cdot \frac{2\alpha^2-\alpha}{2\alpha^2-1} (F(w_{k,t}^{\text{ag},m}) - F^*) \\ &\leq \frac{\frac{\mu}{3}(2\alpha^2-\alpha)}{2\alpha^2-1} L \|w_{k,t}^m - w^*\|^2 + 2L \cdot \frac{2\alpha^2-\alpha}{2\alpha^2-1} (F(w_{k,t}^{\text{ag},m}) - F^*) \\ &= \frac{2\alpha^2-\alpha}{2\alpha^2-1} \cdot 2L\Phi_{k,t}^m \end{aligned} \quad (54)$$

It is easy to show  $(\alpha-1)L \leq \frac{\mu}{3}(2\alpha^2-\alpha)$  by using the fact  $\gamma L \leq 1$ . Therefore, we finally get

$$\begin{aligned} &\mathbb{E}[\|w_{k,t+1}^m - w_{k,t+1}^{\text{ag},m}\|^2 | \mathcal{F}_{k,t}] \\ &\leq (1 - \alpha^{-1})^2 (1 + 2\alpha^{-1}) \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 + (\gamma - \eta)^2 \left( 1 + \frac{\alpha}{2} \right) \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + (\gamma - \eta)^2 \sigma^2 \\ &\leq (1 - \alpha^{-1})^2 (1 + 2\alpha^{-1}) \|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 + (\gamma - \eta)^2 \left( 1 + \frac{\alpha}{2} \right) \left( \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\Phi_{k,t}^m \right) \\ &\quad + (\gamma - \eta)^2 \sigma^2 \end{aligned} \quad (55)$$

Now, let's compute  $\mathbb{E}[\Phi_{k,t+1}^m | \mathcal{F}_{k,t}]$ . We need to compute  $\mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}]$  and  $\mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}]$  first.

$$\begin{aligned}
& \mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}] \\
&= \mathbb{E}[\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma g_{k,t}^m - w^*\|^2 | \mathcal{F}_{k,t}] \\
&\leq \|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma \nabla F(w_{k,t}^{\text{md},m}) - w^*\|^2 + \gamma^2 \sigma^2 \\
&\leq (1 + \frac{1}{2}\alpha^{-1})\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - \gamma \nabla F(w_{k,t}^{\text{md},m}) - w^*\|^2 + \gamma^2 \sigma^2 \\
&= (1 + \frac{1}{2}\alpha^{-1})\|(1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*\|^2 + \gamma^2(1 + \frac{1}{2}\alpha^{-1})\|\nabla F(w_{k,t}^{\text{md},m})\|^2 \\
&\quad - 2\gamma(1 + \frac{1}{2}\alpha^{-1})\langle (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle + \gamma^2 \sigma^2 \\
&\leq (1 + \frac{1}{2}\alpha^{-1})\left(\|(1 - \alpha^{-1})w_{k,t}^m - w^*\|^2 + \alpha^{-1}\|w_{k,t}^{\text{md},m} - w^*\|^2\right) + \gamma^2(1 + \frac{1}{2}\alpha^{-1}) \\
&\quad \cdot \|\nabla F(w_{k,t}^{\text{md},m})\|^2 - 2\gamma(1 + \frac{1}{2}\alpha^{-1})\langle (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle + \gamma^2 \sigma^2
\end{aligned}$$

It is easy to show  $(1 + \frac{1}{2}\alpha^{-1})(1 - \alpha^{-1}) < 1 - \frac{1}{2}\alpha^{-1}$ ,  $1 + \frac{1}{2}\alpha^{-1} \leq \frac{3}{2}$ . Due to these facts, we obtain

$$\begin{aligned}
& \mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}] \\
&\leq (1 - \frac{1}{2}\alpha^{-1})\|w_{k,t}^m - w^*\|^2 + \frac{3}{2}\alpha^{-1}\|w_{k,t}^{\text{md},m} - w^*\|^2 + \frac{3}{2}\gamma^2\|\nabla F(w_{k,t}^{\text{md},m})\|^2 \\
&\quad - 2\gamma(1 + \frac{1}{2}\alpha^{-1})\langle (1 - \alpha^{-1})w_{k,t}^m + \alpha^{-1}w_{k,t}^{\text{md},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle + \gamma^2 \sigma^2 \\
&\leq (1 - \frac{1}{2}\alpha^{-1})\|w_{k,t}^m - w^*\|^2 + \frac{3}{2}\alpha^{-1}\|w_{k,t}^{\text{md},m} - w^*\|^2 + \frac{3}{2}\gamma^2\|\nabla F(w_{k,t}^{\text{md},m})\|^2 \\
&\quad - 2\gamma(1 + \frac{1}{2}\alpha^{-1})\langle (1 - \alpha^{-1}(1 - \beta^{-1}))w_{k,t}^m + \alpha^{-1}(1 - \beta^{-1})w_{k,t}^{\text{ag},m} - w^*, \nabla F(w_{k,t}^{\text{md},m}) \rangle + \gamma^2 \sigma^2
\end{aligned}$$

Next, we compute the upper bound of  $\mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}]$ .

$$\begin{aligned}
& \mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}] \\
&\leq \mathbb{E}[F(w_{k,t}^{\text{md},m}) + \langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t+1}^{\text{ag},m} - w_{k,t}^{\text{md},m} \rangle + \frac{L}{2}\|w_{k,t+1}^{\text{ag},m} - w_{k,t}^{\text{md},m}\|^2 - F^* | \mathcal{F}_{k,t}] \\
&\leq F(w_{k,t}^{\text{md},m}) - F^* - \eta\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&\leq F(w_{k,t}^{\text{md},m}) - F^* - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 (\because 1 - \frac{\eta L}{2} \geq \frac{1}{2} \leftarrow \eta \in [0, \frac{1}{L}]) \\
&= (1 - \frac{1}{2}\alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) + \frac{1}{2}\alpha^{-1}(F(w_{k,t}^{\text{md},m}) - F^*) \\
&\quad + (1 - \frac{1}{2}\alpha^{-1})(F(w_{k,t}^{\text{md},m}) - F(w_{k,t}^{\text{ag},m})) - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&\leq (1 - \frac{1}{2}\alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) - \frac{\mu\alpha^{-1}}{4}\|w_{k,t}^{\text{md},m} - w^*\|^2 + \frac{1}{2}\alpha^{-1}\langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t}^{\text{md},m} - w^* \rangle \\
&\quad + (1 - \frac{1}{2}\alpha^{-1})\langle \nabla F(w_{k,t}^{\text{md},m}), w_{k,t}^{\text{md},m} - w_{k,t}^{\text{ag},m} \rangle - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&= (1 - \frac{1}{2}\alpha^{-1})(F(w_{k,t}^{\text{ag},m}) - F^*) - \frac{\mu\alpha^{-1}}{4}\|w_{k,t}^{\text{md},m} - w^*\|^2 - \frac{\eta}{2}\|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{\eta^2 L}{2}\sigma^2 \\
&\quad + \frac{1}{2}\alpha^{-1}\langle \nabla F(w_{k,t}^{\text{md},m}), 2\alpha\beta^{-1}w_{k,t}^m + (1 - 2\alpha\beta^{-1})w_{k,t}^{\text{ag},m} - w^* \rangle
\end{aligned}$$

It is easy to show  $\frac{1}{2}\alpha^{-1} = \frac{\gamma\mu}{3}(1 + \frac{1}{2}\alpha^{-1})$ . Then, we bound  $\mathbb{E}[\Phi_{k,t+1}^m | \mathcal{F}_{k,t}]$  by using the above results.

$$\begin{aligned}
\mathbb{E}[\Phi_{k,t+1}^m | \mathcal{F}_{k,t}] &= \frac{\mu}{6} \mathbb{E}[\|w_{k,t+1}^m - w^*\|^2 | \mathcal{F}_{k,t}] + \mathbb{E}[F(w_{k,t+1}^{\text{ag},m}) - F^* | \mathcal{F}_{k,t}] \\
&\leq (1 - \frac{1}{2}\alpha^{-1})\Phi_{k,t}^m - \frac{2\eta - \gamma^2\mu}{4} \|\nabla F(w_{k,t}^{\text{md},m})\|^2 + \frac{1}{2}(\frac{\gamma^2\mu}{3} + \eta^2 L)\sigma^2 \\
&\leq (1 - \frac{1}{2}\alpha^{-1})\Phi_{k,t}^m + \frac{1}{2}(\frac{\gamma^2\mu}{3} + \eta^2 L)\sigma^2 \quad (\because \gamma \leq \sqrt{\frac{\eta}{\mu}}) \\
&\leq (1 - \frac{1}{2}\alpha^{-1})\Phi_{k,t}^m + \frac{\gamma^2}{2}(\frac{\mu}{3} + L)\sigma^2
\end{aligned} \tag{56}$$

Plugging (55), (56) in (53) yields,

$$\begin{aligned}
&\mathbb{E}[B_{k,t+1}^m | \mathcal{F}_{k,t}] \\
&\leq \left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right) \left((1 - \alpha^{-1})^2(1 + 2\alpha^{-1})\|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2\right. \\
&\quad \left.+ (\gamma - \eta)^2(1 + \frac{\alpha}{2}) \cdot \left(\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\Phi_{k,t}^m\right) + (\gamma - \eta)^2\sigma^2\right) \\
&\quad \left.+ \gamma^2(\frac{\mu}{3} + L)\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\left((1 - \frac{1}{2}\alpha^{-1})\Phi_{k,t}^m + \frac{\gamma^2}{2}(\frac{\mu}{3} + L)\sigma^2\right)\right) \\
&= (1 - \alpha^{-1})^2(1 + 2\alpha^{-1})\left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right)\|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2 \\
&\quad \left.+ \left(\left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right)(\gamma - \eta)^2(1 + \frac{\alpha}{2}) + (1 - \frac{1}{2}\alpha^{-1})\gamma^2(\frac{\mu}{3} + L)\right)\right. \\
&\quad \left.\cdot \left(\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\Phi_{k,t}^m\right) + \left(\left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right)(\gamma - \eta)^2 + \gamma^4(\frac{\mu}{3} + L)^2\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1}L\right)\sigma^2\right)
\end{aligned} \tag{57}$$

We can show that both coefficients of  $\|w_{k,t}^m - w_{k,t}^{\text{md},m}\|^2$  and  $\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\Phi_{k,t}^m$  are upper bounded by  $1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}$ .

$$\begin{aligned}
(1 - \alpha^{-1})^2(1 + 2\alpha^{-1}) &\leq 1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}} (< 1) \\
\Leftrightarrow 1 - \frac{1}{4}\alpha^{-2} + \frac{1}{2}\alpha^{-1} - (1 - \alpha^{-1})^2(1 + 2\alpha^{-1})(1 + \frac{1}{2}\alpha^{-1}) &\geq 0
\end{aligned} \tag{58}$$

Let's define  $g_1(\alpha^{-1}) = 1 - \frac{1}{4}\alpha^{-2} + \frac{1}{2}\alpha^{-1} - (1 - \alpha^{-1})^2(1 + 2\alpha^{-1})(1 + \frac{1}{2}\alpha^{-1})$ . Then, it is easy to check that  $g_1(\alpha^{-1}) \geq 0$  for  $0 < \alpha^{-1} \leq 1$ . Moreover, we would like to show the below inequality.

$$\begin{aligned}
&\left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right)(\gamma - \eta)^2(1 + \frac{\alpha}{2}) + (1 - \frac{1}{2}\alpha^{-1})\gamma^2(\frac{\mu}{3} + L) \\
&\leq \left(\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2}\right)\gamma^2(1 + \frac{\alpha}{2}) + (1 - \frac{1}{2}\alpha^{-1})\gamma^2(\frac{\mu}{3} + L) \\
&\leq (1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}})\gamma^2(\frac{\mu}{3} + L)
\end{aligned} \tag{59}$$

Since  $\frac{\mu\alpha^{-2}}{3}(1 - \beta^{-1})^2 + L\beta^{-2} = \frac{\mu}{3}(\frac{2\alpha-1}{2\alpha^2-1})^2 + L(\frac{\alpha-1}{2\alpha^2-1})^2 \leq (\frac{\mu}{3} + \frac{L}{4})(\frac{2\alpha-1}{2\alpha^2-1})^2$ , it is enough to show

$$\left(\frac{\mu}{3} + \frac{L}{4}\right)\left(\frac{2\alpha-1}{2\alpha^2-1}\right)^2\gamma^2(1 + \frac{\alpha}{2}) \leq \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\gamma^2(\frac{\mu}{3} + L)$$

We also know that  $\frac{\frac{\mu}{3}+L}{\frac{\mu}{3}+\frac{L}{4}} = 4 - \frac{1}{\frac{1}{3}+\frac{L}{\mu}\frac{1}{4}} > \frac{16}{7}$  ( $\because \frac{L}{\mu} > 1$ ). Then, we only need to show

$$\begin{aligned} \left(\frac{2\alpha-1}{2\alpha^2-1}\right)^2 \left(1 + \frac{\alpha}{2}\right) &\leq \frac{16}{7} \cdot \frac{\frac{1}{2}}{\alpha + \frac{1}{2}} \\ \Leftrightarrow \frac{8}{7}(2\alpha^2-1)^2 - (2\alpha-1)^2 \left(1 + \frac{\alpha}{2}\right) \left(\alpha + \frac{1}{2}\right) &\geq 0 \end{aligned}$$

Let's define  $g_2(\alpha) = \frac{8}{7}(2\alpha^2-1)^2 - (2\alpha-1)^2 \left(1 + \frac{\alpha}{2}\right) \left(\alpha + \frac{1}{2}\right)$ . Then, it is easy to check  $g_2(\alpha) \geq 0$  for  $\alpha \geq \frac{3}{2}$ . As we assume  $\gamma\mu \leq \frac{3}{4}$ , we can say  $\alpha = \frac{3}{2\gamma\mu} - \frac{1}{2} \geq \frac{3}{2}$ . This indicates that the inequality (59) is satisfied. Thus, from (57), (58), and (59) we finally get

$$\begin{aligned} \mathbb{E}[B_{k,t+1}^m | \mathcal{F}_{k,t}] &\leq \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right) B_{k,t}^m \\ &\quad + \left(\left(\frac{\mu\alpha^{-2}}{3}(1-\beta^{-1})^2 + L\beta^{-2}\right)(\gamma-\eta)^2 + \gamma^4\left(\frac{\mu}{3}+L\right)^2 \frac{2\alpha^2-\alpha}{2\alpha^2-1}L\right)\sigma^2 \end{aligned}$$

From this relationship between  $B_{k,t+1}^m$  and  $B_{k,t}^m$ , we obtain the result of Proposition 21.

$$\begin{aligned} \mathbb{E}[B_{k,t}^m] &\leq \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^t \mathbb{E}[B_{k,0}^m] + \left(\left(\frac{\mu\alpha^{-2}}{3}(1-\beta^{-1})^2 + L\beta^{-2}\right)(\gamma-\eta)^2\right. \\ &\quad \left.+ \gamma^4\left(\frac{\mu}{3}+L\right)^2 \frac{2\alpha^2-\alpha}{2\alpha^2-1}L\right)\sigma^2 \cdot \frac{1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^t}{1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)} \\ &\leq \mathbb{E}[B_{k,0}^m] + \left(\left(\frac{\mu}{3}\left(\frac{2\alpha-1}{2\alpha^2-1}\right)^2 + L\left(\frac{\alpha-1}{2\alpha^2-1}\right)^2\right) \cdot (\gamma-\eta)^2 + \gamma^4\left(\frac{\mu}{3}+L\right)^2 \frac{2\alpha^2-\alpha}{2\alpha^2-1}L\right) \\ &\quad \cdot \frac{1 + \frac{1}{2}\alpha^{-1}}{\frac{1}{4}\alpha^{-2}} \cdot \left(1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^t\right)\sigma^2 \end{aligned}$$

**Proposition 22:** Let  $F$  be  $\mu$ -strongly convex, and assume Assumption 2, 3, 4, then for  $\alpha = \frac{3}{2\gamma\mu} - \frac{1}{2}$ ,  $\beta = \frac{2\alpha^2-1}{\alpha-1}$ ,  $\gamma \in [\eta, \sqrt{\frac{\eta}{\mu}}]$ ,  $\eta, \gamma \in (0, \frac{1}{L}]$ ,  $\gamma\mu \leq \frac{3}{4}$ ,  $\tau \geq 2$ , FedAQ yields

$$\begin{aligned} \frac{\mu}{6}\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2}\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] &\leq \left(\gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3}+L\right)\right)\tau^2\mathbb{E}[\Phi_k] \\ &\quad + \left(\frac{\gamma^2\mu}{3} + \eta^2L\right)\tau\sigma^2 \\ &\quad + \left((\gamma-\eta)^2\gamma^2\mu^2\left(\frac{\mu}{3}+\frac{L}{4}\right) + \gamma^4\left(\frac{\mu}{3}+L\right)^2L\right)\frac{\tau^3\sigma^2}{2} \end{aligned}$$

Proof of Proposition 22 We use the same upper bounds for  $\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2]$  and  $\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2]$  as in Proposition 15.

$$\begin{aligned} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] &\leq \tau \left( \sum_{t=0}^{\tau-1} 2\alpha^{-2}(1-\beta^{-1})^2 \mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\gamma^2 \mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2] \right) \\ &\quad + 2\tau\gamma^2\sigma^2 \\ \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] &\leq \tau \left( \sum_{t=0}^{\tau-1} 2\beta^{-2} \mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + 2\eta^2 \mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2] \right) + 2\tau\eta^2\sigma^2 \end{aligned}$$

Thus, by using the above results, we get

$$\begin{aligned}
& \frac{\mu}{6} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2} \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
& \leq \tau \sum_{t=0}^{\tau-1} \left\{ \left( \frac{\mu\alpha^{-2}}{3} (1 - \beta^{-1})^2 + L\beta^{-2} \right) \mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \mathbb{E}[\|\nabla F(w_{k,t}^{\text{md},m})\|^2] \right\} \\
& + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \tau \sigma^2 \\
& \leq \tau \sum_{t=0}^{\tau-1} \left\{ \left( \frac{\mu\alpha^{-2}}{3} (1 - \beta^{-1})^2 + L\beta^{-2} \right) \mathbb{E}[\|w_{k,t}^m - w_{k,t}^{\text{ag},m}\|^2] + \gamma^2 \left( \frac{\mu}{3} + L \right) \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} 2L \mathbb{E}[\Phi_{k,t}^m] \right\} \\
& + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \tau \sigma^2 (\because (54)) \\
& = \tau \left( \sum_{t=0}^{\tau-1} \mathbb{E}[B_{k,t}^m] \right) + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \tau \sigma^2
\end{aligned}$$

By Proposition 21 and the fact  $\Phi_{k,0}^m = \Phi_k$ , we obtain

$$\begin{aligned}
& \frac{\mu}{6} \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2} \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
& \leq \tau \left\{ \sum_{t=0}^{\tau-1} \mathbb{E}[B_{k,0}^m] + \left( \left( \frac{\mu}{3} \left( \frac{2\alpha - 1}{2\alpha^2 - 1} \right)^2 + L \left( \frac{\alpha - 1}{2\alpha^2 - 1} \right)^2 \right) (\gamma - \eta)^2 + \gamma^4 \left( \frac{\mu}{3} + L \right)^2 \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} L \right) \right. \\
& \left. \frac{1 + \frac{1}{2}\alpha^{-1}}{\frac{1}{4}\alpha^{-2}} \left( 1 - \left( 1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}} \right)^t \right) \sigma^2 \right\} + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \tau \sigma^2 \\
& = \tau^2 \left( \left( \frac{\mu\alpha^{-2}}{3} (1 - \beta^{-1})^2 + L\beta^{-2} \right) \mathbb{E}[\|w_k - w_k^{\text{ag}}\|^2] + \gamma^2 \left( \frac{\mu}{3} + L \right) \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L \mathbb{E}[\Phi_k] \right) \\
& + \tau \left( \left( \frac{\mu}{3} \left( \frac{2\alpha - 1}{2\alpha^2 - 1} \right)^2 + L \left( \frac{\alpha - 1}{2\alpha^2 - 1} \right)^2 \right) \cdot (\gamma - \eta)^2 + \gamma^4 \left( \frac{\mu}{3} + L \right)^2 \frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} L \right) \frac{1 + \frac{1}{2}\alpha^{-1}}{\frac{1}{4}\alpha^{-2}} \\
& \cdot \left( \sum_{t=0}^{\tau-1} 1 - \left( 1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}} \right)^t \right) \sigma^2 + \left( \frac{\gamma^2\mu}{3} + \eta^2 L \right) \tau \sigma^2
\end{aligned}$$

Before we get to the final result, let's find the upper bound for  $\|w_k - w_k^{\text{ag}}\|^2$ ,  $\sum_{t=0}^{\tau-1} \left( 1 - \left( 1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}} \right)^t \right)$

$$\begin{aligned}
\|w_k - w_k^{\text{ag}}\|^2 & = \|w_k - w^* - (w_k^{\text{ag}} - w^*)\|^2 \\
& \leq \left( 1 + \frac{1}{3} \right) \|w_k - w^*\|^2 + (1 + 3) \|w_k^{\text{ag}} - w^*\|^2 \\
& \leq \frac{4}{3} \|w_k - w^*\|^2 + 4 \cdot \frac{2}{\mu} \left( F(w_k^{\text{ag}}) - F^* - \langle \nabla F(w^*), w_k^{\text{ag}} - w^* \rangle \right) \\
& = \frac{4}{3} \|w_k - w^*\|^2 + \frac{8}{\mu} (F(w_k^{\text{ag}}) - F^*) = \frac{8}{\mu} \Phi_k
\end{aligned}$$

$$\begin{aligned}
\sum_{t=0}^{\tau-1} \left(1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^t\right) &= \tau - \sum_{t=0}^{\tau-1} \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^t \\
&= \tau - \frac{1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)^\tau}{1 - \left(1 - \frac{1}{2}\alpha^{-1} + \frac{\frac{1}{2}\alpha^{-1}}{1 + \frac{1}{2}\alpha^{-1}}\right)} \\
&\leq \tau - \frac{1 - \left(1 - \frac{\frac{1}{4}\alpha^{-2}}{1 + \frac{1}{2}\alpha^{-1}}\tau + \left(\frac{\frac{1}{4}\alpha^{-2}}{1 + \frac{1}{2}\alpha^{-1}}\right)^2 \frac{\tau(\tau-1)}{2}\right)}{\frac{\frac{1}{4}\alpha^{-2}}{1 + \frac{1}{2}\alpha^{-1}}} \\
&= \frac{\frac{1}{4}\alpha^{-2}}{1 + \frac{1}{2}\alpha^{-1}} \cdot \frac{\tau(\tau-1)}{2} \leq \frac{\frac{1}{4}\alpha^{-2}}{1 + \frac{1}{2}\alpha^{-1}} \cdot \frac{\tau^2}{2}
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\frac{\mu}{6}\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2}\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\
&\leq \left(\frac{8}{3}\alpha^{-2}(1 - \beta^{-1})^2 + \frac{8L}{\mu}\beta^{-2} + \gamma^2\left(\frac{\mu}{3} + L\right)\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L\right)\tau^2\mathbb{E}[\Phi_k] + \left(\frac{\gamma^2\mu}{3} + \eta^2L\right)\tau\sigma^2 \\
&+ \left(\left(\frac{\mu}{3}\left(\frac{2\alpha - 1}{2\alpha^2 - 1}\right)^2 + L\left(\frac{\alpha - 1}{2\alpha^2 - 1}\right)^2\right) \cdot (\gamma - \eta)^2 + \gamma^4\left(\frac{\mu}{3} + L\right)^2\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1}L\right) \cdot \frac{\tau^3\sigma^2}{2} \tag{60}
\end{aligned}$$

Moreover, we can simplify the above inequality by replacing  $\alpha, \beta$  with  $\gamma, \mu$ . It is easy to show  $\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \leq 1$ ,  $\frac{2\alpha - 1}{2\alpha^2 - 1} \leq \frac{1}{\alpha} = \frac{2\gamma\mu}{3 - \gamma\mu} \leq \gamma\mu$ . Then, we can further show

$$\begin{aligned}
&\frac{8}{3}\alpha^{-2}(1 - \beta^{-1})^2 + \frac{8L}{\mu}\beta^{-2} + \gamma^2\left(\frac{\mu}{3} + L\right)\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L \\
&= \frac{8}{3}\left(\frac{2\alpha - 1}{2\alpha^2 - 1}\right)^2 + \frac{8L}{\mu}\left(\frac{\alpha - 1}{2\alpha^2 - 1}\right)^2 + \gamma^2\left(\frac{\mu}{3} + L\right)\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1} \cdot 2L \\
&\leq \left(\frac{8}{3} + \frac{2L}{\mu}\right)\left(\frac{2\alpha - 1}{2\alpha^2 - 1}\right)^2 + \gamma^2\left(\frac{\mu}{3} + L\right)2L \\
&\leq \left(\frac{8}{3} + \frac{2L}{\mu}\right)\alpha^{-2} + \gamma^2\left(\frac{\mu}{3} + L\right)2L \\
&\leq \gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3} + L\right) \tag{61}
\end{aligned}$$

We also get

$$\begin{aligned}
&\left(\frac{\mu}{3}\left(\frac{2\alpha - 1}{2\alpha^2 - 1}\right)^2 + L\left(\frac{\alpha - 1}{2\alpha^2 - 1}\right)^2\right) \cdot (\gamma - \eta)^2 + \gamma^4\left(\frac{\mu}{3} + L\right)^2\frac{2\alpha^2 - \alpha}{2\alpha^2 - 1}L \\
&\leq \left(\frac{\mu}{3} + \frac{L}{4}\right)\left(\frac{2\alpha - 1}{2\alpha^2 - 1}\right)^2(\gamma - \eta)^2 + \gamma^4\left(\frac{\mu}{3} + L\right)^2L \\
&\leq (\gamma - \eta)^2\gamma^2\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + \gamma^4\left(\frac{\mu}{3} + L\right)^2L \tag{62}
\end{aligned}$$

Finally, from (60), (61), and (62), we conclude as below

$$\begin{aligned} \frac{\mu}{6}\mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{L}{2}\mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] &\leq \left(\gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3} + L\right)\right)\tau^2\mathbb{E}[\Phi_k] \\ &\quad + \left(\frac{\gamma^2\mu}{3} + \eta^2L\right)\tau\sigma^2 \\ &\quad + \left((\gamma - \eta)^2\gamma^2\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + \gamma^4\left(\frac{\mu}{3} + L\right)^2L\right)\frac{\tau^3\sigma^2}{2} \end{aligned}$$

Proof of Lemma 5 By the definition of  $\Phi_k$ ,  $\Phi_{k,t}$  and Proposition 18,

$$\mathbb{E}[\Phi_{k+1}] = \mathbb{E}[\Phi_{k,\tau}] + \frac{\mu}{6}\mathbb{E}[\|w_{k+1} - \bar{w}_{k,\tau}\|^2] + \mathbb{E}[F(w_{k+1}^{\text{ag}}) - F(\bar{w}_{k,\tau}^{\text{ag}})]$$

Applying Proposition 19 and Proposition 20, we have

$$\begin{aligned} &\mathbb{E}[\Phi_{k+1}] \\ &\leq \left(1 - \frac{1}{3}\gamma\mu\right)^\tau\mathbb{E}[\Phi_k] + \left(\frac{\eta^2L}{2} + \frac{\gamma^2\mu}{6}\right)\frac{\tau\sigma^2}{M} + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ &\quad + \frac{q\mu}{6M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^m - w_k\|^2] + \frac{qL}{2M^2} \sum_{m=1}^M \mathbb{E}[\|w_{k,\tau}^{\text{ag},m} - w_k^{\text{ag}}\|^2] \\ &\leq \left(1 - \frac{1}{3}\gamma\mu\right)^\tau\mathbb{E}[\Phi_k] + \left(\frac{\eta^2L}{2} + \frac{\gamma^2\mu}{6}\right)\frac{\tau\sigma^2}{M} + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ &\quad + \frac{q}{M} \left[ \left(\gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3} + L\right)\right)\tau^2\mathbb{E}[\Phi_k] + \left(\frac{\gamma^2\mu}{3} + \eta^2L\right)\tau\sigma^2 \right. \\ &\quad \left. + \left((\gamma - \eta)^2\gamma^2\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + \gamma^4\left(\frac{\mu}{3} + L\right)^2L\right)\frac{\tau^3\sigma^2}{2} \right] \\ &= D(\gamma, \tau)\mathbb{E}[\Phi_k] + \left(\frac{\eta^2L}{2} + \frac{\gamma^2\mu}{6}\right)\frac{\tau\sigma^2}{M} + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ &\quad + \frac{q}{M} \left(\frac{\gamma^2\mu}{3} + \eta^2L\right)\tau\sigma^2 + \frac{q}{2M} \left((\gamma - \eta)^2\gamma^2\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + \gamma^4\left(\frac{\mu}{3} + L\right)^2L\right)\tau^3\sigma^2 \end{aligned}$$

The second inequality comes from Proposition 22.  $D(\gamma, \tau)$  is defined as below.

$$D(\gamma, \tau) = \left(1 - \frac{1}{3}\gamma\mu\right)^\tau + \frac{q}{M} \left(\gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3} + L\right)\right)\tau^2$$

#### 5.4.2 Proof of Theorem 6

Proof of Theorem 6 At first, due to the condition (36) in Theorem 6, we get

$$\begin{aligned} D(\gamma, \tau) &= \left(1 - \frac{1}{3}\gamma\mu\right)^\tau + \frac{q}{M} \left(\gamma^2\mu\left(\frac{8}{3}\mu + 2L\right) + 2\gamma^2L\left(\frac{\mu}{3} + L\right)\right)\tau^2 \\ &\leq 1 - \frac{1}{3}\gamma\mu\tau + \frac{1}{9}\gamma^2\mu^2\tau^2 + \frac{q}{M}\gamma^2 \left(\mu\left(\frac{8}{3}\mu + 2L\right) + 2L\left(\frac{\mu}{3} + L\right)\right)\tau^2 \\ &= 1 - \frac{1}{3}\gamma\mu\tau + \left(\frac{1}{9}\mu^2 + \frac{q}{M} \left(\mu\left(\frac{8}{3}\mu + 2L\right) + 2L\left(\frac{\mu}{3} + L\right)\right)\right)\gamma^2\tau^2 \\ &\leq 1 - \frac{1}{6}\gamma\mu\tau \quad (\because \text{condition (36)}) \end{aligned}$$

It is trivial that  $\gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}}) \in [\eta, \sqrt{\frac{\eta}{\mu}}]$ . Thus, we can use Lemma 5. By using Lemma 5 and the above result, we obtain

$$\begin{aligned} & \mathbb{E}[\Phi_{k+1}] \\ & \leq (1 - \frac{1}{6}\gamma\mu\tau)\mathbb{E}[\Phi_k] + (\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6})\frac{\tau\sigma^2}{M} + \gamma\tau \cdot \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \\ & \quad + \frac{q}{M}(\frac{\gamma^2 \mu}{3} + \eta^2 L)\tau\sigma^2 + \frac{q}{2M} \left( (\gamma - \eta)^2 \gamma^2 \mu^2 (\frac{\mu}{3} + \frac{L}{4}) + \gamma^4 (\frac{\mu}{3} + L)^2 L \right) \tau^3 \sigma^2 \end{aligned} \quad (63)$$

By the Lemma C.14 in [43], we know that the below quantity is bounded.

$$\begin{aligned} & \max_{0 \leq t < \tau} \mathbb{E}[\|\nabla F(\bar{w}_{k,t}^{\text{md}}) - \frac{1}{M} \sum_{m=1}^M \nabla F(w_{k,t}^{\text{md},m})\|^2] \leq B' \\ & B' = \begin{cases} 4\eta^2 L^2 \tau \sigma^2 \left(1 + \frac{\gamma^2 \mu}{\eta}\right)^{2\tau}, & \text{if } \gamma \in \left(\eta, \sqrt{\frac{\eta}{\mu}}\right) \\ 4\eta^2 L^2 \tau \sigma^2, & \text{if } \gamma = \eta \end{cases} \end{aligned}$$

Telescoping (63) yields

$$\begin{aligned} \mathbb{E}[\Phi_K] & \leq (1 - \frac{1}{6}\gamma\mu\tau)^K \Phi_0 + \left( \sum_{k'=0}^{K-1} (1 - \frac{1}{6}\gamma\mu\tau)^{k'} \right) \cdot \left[ (\frac{\eta^2 L}{2} + \frac{\gamma^2 \mu}{6})\frac{\tau\sigma^2}{M} + \frac{q}{M}(\frac{\gamma^2 \mu}{3} + \eta^2 L)\tau\sigma^2 \right. \\ & \quad \left. + \frac{q}{2M} \left( (\gamma - \eta)^2 \gamma^2 \mu^2 (\frac{\mu}{3} + \frac{L}{4}) + \gamma^4 (\frac{\mu}{3} + L)^2 L \right) \tau^3 \sigma^2 + \gamma\tau B' \right] \\ & \leq \exp\left(-\frac{\gamma\mu\tau K}{6}\right) \Phi_0 + \frac{3\eta^2 L \sigma^2}{\gamma\mu M} + \frac{\gamma\sigma^2}{M} + \frac{6B'}{\mu} + 2q \left( \frac{\gamma\sigma^2}{M} + \frac{3\eta^2 L \sigma^2}{\gamma\mu M} \right) \\ & \quad + \frac{3q}{M} \left( (\gamma - \eta)^2 \gamma \mu (\frac{\mu}{3} + \frac{L}{4}) + \frac{\gamma^3 (\frac{\mu}{3} + L)^2 L}{\mu} \right) \tau^2 \sigma^2 \end{aligned}$$

The last inequality comes from the fact that  $\sum_{k'=0}^{K-1} (1 - \frac{1}{6}\gamma\mu\tau)^{k'} \leq \frac{6}{\gamma\mu\tau}$ . Since we plug in  $\gamma = \max(\eta, \sqrt{\frac{\eta}{\mu\tau}})$ , we can use Lemma C.15 in [43]. Therefore, we obtain

$$\begin{aligned} \mathbb{E}[\Phi_K] & \leq \exp\left(-\frac{1}{6} \max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}}) K\tau\right) \Phi_0 + \frac{2(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}} M \tau^{\frac{1}{2}}} + \frac{4(2q+1)\eta^2 L^2 \tau \sigma^2}{\mu} \\ & \quad + \frac{24e^2 \eta^2 L^2 \tau \sigma^2}{\mu} + \frac{3q\tau^2 \sigma^2}{M} \max\left(\frac{\eta^{\frac{3}{2}} \mu (\frac{\mu}{3} + \frac{L}{4})}{\mu^{\frac{3}{2}} \tau^{\frac{3}{2}}}, \frac{\eta^{\frac{3}{2}} (\frac{\mu}{3} + L)^2 L}{\mu^{\frac{5}{2}} \tau^{\frac{3}{2}}}, \frac{\eta^3 (\frac{\mu}{3} + L)^2 L}{\mu}\right) \end{aligned}$$

The first term stems directly from Lemma C.15 in [43]. Also, the last term comes from the fact that

$$(\gamma - \eta)^2 \gamma \mu (\frac{\mu}{3} + \frac{L}{4}) + \frac{\gamma^3 (\frac{\mu}{3} + L)^2 L}{\mu} \leq \begin{cases} \gamma^3 \mu (\frac{\mu}{3} + \frac{L}{4}) + \frac{\gamma^3 (\frac{\mu}{3} + L)^2 L}{\mu}, & \text{if } \gamma \neq \eta \\ \frac{\eta^3 (\frac{\mu}{3} + L)^2 L}{\mu}, & \text{if } \gamma = \eta \end{cases}$$

Therefore, by simple inequalities such as  $\max(a, b) \leq a + b$  and  $\min(a, b) \leq a$ , we ultimately get

$$\begin{aligned} \mathbb{E}[\Phi_K] & \leq \exp\left(-\frac{1}{6} \max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}}) K\tau\right) \Phi_0 + \frac{2(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}} M \tau^{\frac{1}{2}}} + \frac{8(q+25)\eta^2 L^2 \tau \sigma^2}{\mu} \\ & \quad + \frac{3q \left( \mu^2 (\frac{\mu}{3} + \frac{L}{4}) + L (\frac{\mu}{3} + L)^2 \right) \eta^{\frac{3}{2}} \tau^{\frac{1}{2}} \sigma^2}{\mu^{\frac{5}{2}} M} + \frac{3qL (\frac{\mu}{3} + L)^2 \eta^3 \tau^2 \sigma^2}{\mu M} \end{aligned} \quad (64)$$



### 5.4.3 Proof of Corollary 23

**Corollary 23:** Let  $D_1, D_2$ , and  $\eta_0$  as below. Note that  $T = K\tau$ .

$$D_1 = \frac{\left(\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + L\left(\frac{\mu}{3} + L\right)^2\right)q}{\mu^{\frac{5}{2}}}, \quad D_2 = \frac{q\left(\frac{\mu}{3} + L\right)^2L}{\mu}$$

$$\eta_0 = \frac{36\tau}{\mu T^2} \log^2 \left( e + \min\left(\frac{\mu MT \Phi_0}{(2q+1)\sigma^2}, \frac{\mu^3 T^4 \Phi_0}{(q+25)L^2 \tau^3 \sigma^2}, \frac{\mu^3 MT^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2}\right) \right)$$

Then for  $\eta = \min\left(\frac{1}{L}, \eta_0\right)$ , FedAQ yields

$$\mathbb{E}[\Phi_K] \leq \min\left(\exp\left(-\frac{\mu T}{6L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}} T}{6L^{\frac{1}{2}} \tau^{\frac{1}{2}}}\right)\right) \Phi_0$$

$$+ \frac{13(2q+1)\sigma^2}{\mu MT} \log^2 \left( e + \frac{\mu MT \Phi_0}{(2q+1)\sigma^2} \right) \quad (65)$$

$$+ \frac{10369(q+25)L^2 \tau^3 \sigma^2}{\mu^3 T^4} \log^4 \left( e + \frac{\mu^3 T^4 \Phi_0}{(q+25)L^2 \tau^3 \sigma^2} \right) \quad (66)$$

$$+ \frac{649(\mu^{\frac{3}{2}} D_1 + 216 D_2) \tau^2 \sigma^2}{\mu^3 MT^3} \log^6 \left( e + \frac{\mu^3 MT^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 216 D_2) \tau^2 \sigma^2} \right) \quad (67)$$

Proof of Corollary 23 Let's decompose the final result (64) of the Theorem 6 into a decreasing term and an increasing term. We denote the decreasing term  $\phi_1$  and the increasing term  $\phi_2$  as below.

$$\phi_1(\eta) = \exp\left(-\frac{1}{6} \max(\eta\mu, \sqrt{\frac{\eta\mu}{\tau}})T\right) \Phi_0$$

$$\phi_2(\eta) = \frac{2(2q+1)\eta^{\frac{1}{2}}\sigma^2}{\mu^{\frac{1}{2}} M \tau^{\frac{1}{2}}} + \frac{8(q+25)\eta^2 L^2 \tau \sigma^2}{\mu} + \frac{3q\left(\mu^2\left(\frac{\mu}{3} + \frac{L}{4}\right) + L\left(\frac{\mu}{3} + L\right)^2\right)\eta^{\frac{3}{2}}\tau^{\frac{1}{2}}\sigma^2}{\mu^{\frac{5}{2}} M}$$

$$+ \frac{3qL\left(\frac{\mu}{3} + L\right)^2\eta^3\tau^2\sigma^2}{\mu M}$$

Since  $\phi_1$  is the decreasing term, we have

$$\phi_1(\eta) \leq \phi_1\left(\frac{1}{L}\right) + \phi_1(\eta_0) \quad (68)$$

where

$$\phi_1\left(\frac{1}{L}\right) = \min\left(\exp\left(-\frac{\mu T}{6L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}} T}{6L^{\frac{1}{2}} \tau^{\frac{1}{2}}}\right)\right) \Phi_0$$

$$\phi_1(\eta_0) \leq \exp\left(-\frac{1}{6} \sqrt{\frac{\eta_0 \mu}{\tau}} T\right)$$

$$= \left( e + \min\left(\frac{\mu MT \Phi_0}{(2q+1)\sigma^2}, \frac{\mu^3 T^4 \Phi_0}{(q+25)L^2 \tau^3 \sigma^2}, \frac{\mu^3 MT^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2}\right) \right)^{-1} \Phi_0$$

$$\leq \frac{(2q+1)\sigma^2}{\mu MT} + \frac{(q+25)L^2 \tau^3 \sigma^2}{\mu^3 T^4} + \frac{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2}{\mu^3 MT^3}$$

Since  $\phi_2$  is the increasing term, we have

$$\begin{aligned}
& \phi_2(\eta) \\
& \leq \phi_2(\eta_0) \\
& \leq \frac{12(2q+1)\sigma^2}{\mu MT} \log \left( e + \frac{\mu MT \Phi_0}{(2q+1)\sigma^2} \right) + \frac{8 \cdot 36^2 (q+25) L^2 \tau^3 \sigma^2}{\mu^3 T^4} \log^4 \left( e + \frac{\mu^3 T^4 \Phi_0}{(q+25) L^2 \tau^3 \sigma^2} \right) \\
& + \frac{3 \cdot 6^3 D_1 \tau^2 \sigma^2}{\mu^{\frac{3}{2}} M T^3} \log^3 \left( e + \frac{\mu^3 M T^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2} \right) \\
& + \frac{3 \cdot 36^3 D_2 \tau^5 \sigma^2}{\mu^3 M T^6} \log^6 \left( e + \frac{\mu^3 M T^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2} \right) \\
& \leq \frac{12(2q+1)\sigma^2}{\mu MT} \log \left( e + \frac{\mu MT \Phi_0}{(2q+1)\sigma^2} \right) + \frac{8 \cdot 36^2 (q+25) L^2 \tau^3 \sigma^2}{\mu^3 T^4} \log^4 \left( e + \frac{\mu^3 T^4 \Phi_0}{(q+25) L^2 \tau^3 \sigma^2} \right) \\
& + \frac{3 \cdot 6^3 (\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2}{\mu^3 M T^3} \log^6 \left( e + \frac{\mu^3 M T^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 6^3 D_2) \tau^2 \sigma^2} \right) \tag{69}
\end{aligned}$$

The last inequality comes from  $\frac{\tau}{T} \leq 1$ . Therefore, by combining (68) and (69), we finally get

$$\begin{aligned}
\mathbb{E}[\Phi_K] & \leq \phi_1(\eta) + \phi_2(\eta) \\
& \leq \phi_1\left(\frac{1}{L}\right) + \phi_1(\eta_0) + \phi_2(\eta_0) \\
& \leq \min \left( \exp\left(-\frac{\mu T}{6L}\right), \exp\left(-\frac{\mu^{\frac{1}{2}} T}{6L^{\frac{1}{2}} \tau^{\frac{1}{2}}}\right) \right) \Phi_0 + \frac{13(2q+1)\sigma^2}{\mu MT} \log^2 \left( e + \frac{\mu MT \Phi_0}{(2q+1)\sigma^2} \right) \\
& + \frac{10369(q+25)L^2 \tau^3 \sigma^2}{\mu^3 T^4} \log^4 \left( e + \frac{\mu^3 T^4 \Phi_0}{(q+25)L^2 \tau^3 \sigma^2} \right) \\
& + \frac{649(\mu^{\frac{3}{2}} D_1 + 216 D_2) \tau^2 \sigma^2}{\mu^3 M T^3} \log^6 \left( e + \frac{\mu^3 M T^3 \Phi_0}{(\mu^{\frac{3}{2}} D_1 + 216 D_2) \tau^2 \sigma^2} \right)
\end{aligned}$$

## 5.5 More Theoretical Details about Remark 9 and Contribution 2 in Introduction

### 5.5.1 Why Haddadpour et al. (2021) Cannot Achieve a Linear Speedup

It is hard to say that [9] achieves a linear speedup in  $M$  in strongly-convex and homogeneous settings. Let's first recap Corollary D.8 in [9]. They let  $\eta\gamma\mu\tau \leq \frac{1}{2}$ ,  $\kappa = \frac{L}{\mu}$ ,  $\gamma \geq M$  and tune  $\eta$  as  $\eta = \frac{1}{2L(\frac{q}{M}+1)\tau\gamma}$ . Here,  $\eta$  is the client learning rate, and  $\gamma$  is the server learning rate. Other parameters are the same as we defined. Then, they obtain the below result.

$$\begin{aligned}
\mathbb{E}[F(w_K) - F^*] & \leq \exp(-\eta\gamma\mu\tau K)(F(w_0) - F^*) + \frac{1}{\mu} \left[ \frac{1}{2} \tau L^2 \eta^2 \sigma^2 + (1+q) \frac{\gamma \eta L \sigma^2}{2M} \right] \tag{70} \\
& \leq \mathcal{O} \left( \exp\left(-\frac{K}{2(\frac{q}{M}+1)\kappa}\right) (F(w_0) - F^*) + \frac{\sigma^2}{\gamma^2 \mu \tau} + \frac{(q+1)\sigma^2}{\mu(\frac{q}{M}+1)\tau M} \right) \\
& = \mathcal{O} \left( \exp\left(-\frac{K}{2(\frac{q}{M}+1)\kappa}\right) (F(w_0) - F^*) + \frac{\sigma^2 K}{\gamma^2 \mu T} + \frac{(q+1)K\sigma^2}{\mu(\frac{q}{M}+1)TM} \right)
\end{aligned}$$

Let's focus on the second and third term. We assume  $M$  is large enough and represent them only with  $\gamma, K, T, M$  to easily check the linear speedup of this convergence rate. Then, we obtain

$$\mathcal{O} \left( \frac{K}{\gamma^2 T} + \frac{K}{MT} \right) \leq \mathcal{O} \left( \frac{K}{M^2 T} + \frac{K}{MT} \right) (\because \gamma \geq M) \tag{71}$$

Thus, it seemingly achieves a linear speedup in  $M$  when  $K$  is just a constant. However, we are missing the critical point in this analysis. To be specific, let's consider the case when  $\gamma = 1$ . Then, the convergence rate (71) changes into  $\mathcal{O}\left(\frac{K}{T} + \frac{K}{MT}\right)$  that cannot achieve a linear speedup in  $M$ . This is implausible because the convergence rate (70) becomes tighter when  $\gamma = 1$  than  $\gamma \geq M$  (See the last term of (70)). Actually, we can achieve a linear speedup in  $M$  when  $\gamma = 1$  if we tune  $\eta = \frac{1}{2L(\frac{q}{M}+1)\tau M}$ . However, this is not an appropriate tuning because there is  $M$  in the denominator. Similarly, [9] tunes  $\eta = \frac{1}{2L(\frac{q}{M}+1)\tau\gamma}$  where  $\gamma \geq M$ . Even though there is no  $M$  in the denominator, the condition  $\gamma \geq M$  forcibly makes the convergence rate achieve a linear speedup without any theoretical benefits of the algorithm. Therefore, we cannot say their  $\eta$  makes their algorithm achieve a linear speedup in  $M$ . We should tune in a different way that does not contain  $M$  in a denominator. For reference, our tuning parameter  $\eta$  for the FedAQ algorithm does not contain  $M$  in the denominator (See Corollary 17 and Corollary 23).

### 5.5.2 New Convergence Rate for Haddadpour et al. (2021)

We propose new  $\eta$  and convergence rate for [9]. This new  $\eta$  makes the algorithm achieve a linear speedup in  $M$ . Let's denote  $\Phi_0 = F(w_0) - F^*$ . We also know that  $T = K\tau$ . Then, we choose  $\eta$  as

$$\eta = \frac{1}{\gamma\mu T} \log \left( e + \min\left(\frac{\gamma^2\mu^3T^2\Phi_0}{\tau L^2\sigma^2}, \frac{\mu^2MT\Phi_0}{(1+q)L\sigma^2}\right) \right)$$

We plug in this  $\eta$  to (70). We bound the first term as below.

$$\begin{aligned} \exp(-\eta\gamma\mu\tau K)(F(w_0) - F^*) &= \left( e + \min\left(\frac{\gamma^2\mu^3T^2\Phi_0}{\tau L^2\sigma^2}, \frac{\mu^2MT\Phi_0}{(1+q)L\sigma^2}\right) \right)^{-1} \Phi_0 \\ &\leq \frac{\tau L^2\sigma^2}{\gamma^2\mu^3T^2} + \frac{(1+q)L\sigma^2}{\mu^2MT} \end{aligned}$$

The another terms are bounded as below.

$$\begin{aligned} \frac{1}{\mu} \left[ \frac{1}{2}\tau L^2\eta^2\sigma^2 + (1+q)\frac{\gamma\eta L\sigma^2}{2M} \right] &\leq \frac{\tau L^2\sigma^2}{2\gamma^2\mu^3T^2} \log^2 \left( e + \frac{\gamma^2\mu^3T^2\Phi_0}{\tau L^2\sigma^2} \right) \\ &\quad + \frac{(1+q)L\sigma^2}{2\mu^2MT} \log \left( e + \frac{\mu^2MT\Phi_0}{(1+q)L\sigma^2} \right) \end{aligned}$$

Thus, we obtain a new convergence rate by combining the above two bounds.

$$\begin{aligned} \mathbb{E}[F(w_K) - F^*] &\leq \exp(-\eta\gamma\mu\tau K)(F(w_0) - F^*) + \frac{1}{\mu} \left[ \frac{1}{2}\tau L^2\eta^2\sigma^2 + (1+q)\frac{\gamma\eta L\sigma^2}{2M} \right] \\ &\leq \frac{3\tau L^2\sigma^2}{2\gamma^2\mu^3T^2} \log^2 \left( e + \frac{\gamma^2\mu^3T^2\Phi_0}{\tau L^2\sigma^2} \right) + \frac{3(1+q)L\sigma^2}{2\mu^2MT} \log \left( e + \frac{\mu^2MT\Phi_0}{(1+q)L\sigma^2} \right) \end{aligned}$$

Here, we replace  $\tau$  with  $\frac{T}{K}$ . Then, we represent the above convergence rate with only  $T, K, M, q$ .

$$\tilde{\mathcal{O}}\left(\frac{1}{TK} + \frac{1+q}{MT}\right)$$

This is the new convergence rate we propose in Remark 9. We also get  $K = \tilde{\mathcal{O}}\left(\frac{M}{1+q}\right)$  communication rounds make this algorithm achieve a linear speedup in  $M$ .

### 5.5.3 More Details on Contribution 2 in Introduction

**More Details on  $d_{\text{quant}}$**  This paragraph explains why FedAQ needs to send only  $d_{\text{quant}} = O(\log \frac{1}{q})$  bits for each value. We use the result of Lemma 3.1 in [1]. They show the below result with a low-precision quantizer (Example 1 in section 3)

$$\mathbb{E}[\|Q(x, s) - x\|_2^2] \leq \min\left(\frac{n}{s^2}, \frac{\sqrt{n}}{s}\right)\|x\|_2^2$$

where  $n$  is the dimension of  $x$ , and  $s$  is the number of quantization levels. Then, we regard  $q$  as

$$q = \frac{\sqrt{n}}{s} = \frac{\sqrt{n}}{2^{d_{\text{quant}}}} \quad (72)$$

Thus, we obtain the following conclusion.

$$d_{\text{quant}} = \frac{\frac{1}{2} \log n + \log \frac{1}{q}}{\log 2} = O(\log \frac{1}{q})$$

**Comparing FedAQ to FedAC** We compare computation and communication efficiency of FedAC-II and FedAQ under the condition set (35) to achieve the same error. Let's recall the convergence rate of FedAC and FedAQ. The convergence rate of FedAC and FedAQ is respectively  $\tilde{O}(\frac{1}{MT} + \frac{1}{TK^3})$  and  $\tilde{O}(\frac{1+q}{MT} + \frac{1+q}{TK^3})$ . Let's say FedAC requires  $T$  iterations and  $K = M^{\frac{1}{3}}$  communication rounds to achieve the error  $\frac{1}{MT}$ . Then, FedAQ requires

$$T' = (1 + q)T, \quad K' = M^{\frac{1}{3}}$$

to achieve the same error  $\frac{1}{MT}$ . This means FedAQ needs  $1 + q$  times more local steps and the same number of communication rounds to achieve the same error of FedAC. These local steps do not require any communication with the server hence can be performed without any additional communication overhead.

From discussion in the previous section, if we use the simple low-precision quantizer, we need only  $d_{\text{quant}} = O(\log \frac{1}{q})$  bits for communicating values with enough precision that can lead to an error rate of  $O(\frac{1}{MT})$ . In comparison, FedAC would require  $O(\log(MT))$  bits to maintain enough precision to achieve the same error rate. In a majority of tasks in the real world, 32 bits are usually enough for  $d_{\text{full}}$  to achieve enough precision as we usually don't need converge to a very small error rate. Nonetheless, even if we compare FedAQ(8bits) with to FedAC(32bits), we argue that the overall benefit from less communication by quantization is more influential than the slowdown effect from quantization.

For example, if we consider a  $l_2$ -regularized logistic regression model for MNIST (strongly convex experiment) and quantize from 32 bits to  $d_{\text{quant}} = 8$  bits. Here,  $n = 784 \times 10$ . We get the following results by using (72).

$$1 + q = 1 + \frac{\sqrt{n}}{2^{d_{\text{quant}}}} = 1 + \frac{\sqrt{7840}}{2^8} \simeq 1.346,$$

On the other hand, the ratio of data communicated by FedAC and FedAQ is

$$\frac{32}{d_{\text{quant}}} = 4$$

In contribution 2, we claim  $1 + q \ll \frac{d_{\text{full}}}{d_{\text{quant}}}$  because  $d_{\text{full}}$  is unbounded as  $T$  goes to infinity. In the real world example,  $\frac{d_{\text{full}}}{d_{\text{quant}}} = 4$  is still much greater than  $1 + q$ . Furthermore, since the local computation is much cheaper than data communication, we conclude that the benefit from less communication by quantization (4 times less bits) overwhelm the slowdown effect from quantization ( $(1 + q)$  times more local computation).

## 6 Experiments

In this section, we provide experimental results of FedAQ in homogeneous local data distribution settings. We compare FedAQ with other quantization-based federated optimization algorithms, FedPAQ [27] and FedCOMGATE [9]. FedAvg [26] and FedAC [43], federated optimization algorithms without quantization, are also our baselines. We empirically validate the performance of 5 algorithms on classical classification tasks on MNIST[18] and CIFAR-10[17] datasets in the distributed learning environment. We consider three objective functions i) A strongly convex objective of  $l_2$ -regularized logistic regression model on the MNIST dataset, ii) A non convex objective of training a multilayer perceptron on the MNIST data, and iii) A non convex objective of training a convolution neural network (CNN) on the CIFAR-10 dataset.

### 6.1 Experimental Setup

**Implementation Environment.** We follow the implementation setup in [9]. We use the Distributed library of PyTorch to implement our algorithm because this library allows us to simulate real-world communication and distributed training. The 18 cores of Intel Xeon E5-2676 CPU are used as computing sources. Each core is considered as one local client. We use 16 cores for strongly convex MNIST, 18 cores for the non-convex MNIST, and 8 cores for the CIFAR-10. For MNIST, the strongly convex experiment and the non-convex one respectively run for 300 rounds of communication with 20 local updates and 50 rounds of communication with 100 local updates. The CIFAR-10 experiment runs for 100 rounds of communication with 100 local updates.

**Datasets.** For image classification tasks, we choose two main classical image datasets: MNIST and CIFAR-10. Since we assume homogeneous settings, data is distributed homogeneously among clients, which also means each device has access to all 10 classes.

**Hyperparameter Choice.** The important hyperparameters in our experiments are learning rates for each algorithm. For the client learning rate  $\eta$ , we respectively use 0.002, 0.1, and 0.01 for strongly convex MNIST, non-convex MNIST, and CIFAR-10 for all algorithms. For FedAQ and FedAC, once we set the value of  $\mu$ , other hyperparameters ( $\gamma, \alpha, \beta$ ) are automatically determined (See condition set (34) and (35)). Thus, we choose 0.1, 0.01, and 0.2 for  $\mu$  value for strongly convex MNIST, non-convex MNIST, and CIFAR-10. Since too large  $\mu$  leads to slow convergence and too small  $\mu$  leads to unstable training, we get these  $\mu$  values by tuning  $\mu$  appropriately. FedCOMGATE has a server learning rate, and we set this value as 1 for all experiments.

**Quantization Bits.** We have three quantization-based federated algorithms: FedAQ, FedPAQ, FedCOMGATE. We quantize the updates from 32 bits to 8 bits for all quantization-based algorithms in both MNIST and CIFAR-10. Additionally, particularly for FedAQ in non-convex experiments, we consider 4 bits quantization as well. Since FedAQ sends twice as many messages as FedPAQ or FedCOMGATE at every synchronization when we use 8 bits quantization for all cases, we apply 4 bits quantization to FedAQ to let FedAQ send the same amount of information in each communication round as other quantization-based algorithms for a fair comparison.

**New Time Metric.** In our experiments, communication between CPU cores is very fast, so it is hard to say that the environment of our experiments fully reflects the real-world federated learning when there is a heavy communication burden. Thus, we use a linear model to estimate the execution time  $T_{\text{round}}(\mathcal{A})$  between two

consecutive communication rounds for real federated learning scenarios [35].

$$\begin{aligned}
 T_{\text{round}}(\mathcal{A}) &= T_{\text{comm}}(\mathcal{A}) + T_{\text{comp}}(\mathcal{A}), & T_{\text{comm}}(\mathcal{A}) &= \frac{S_{\text{down}}(\mathcal{A})}{B_{\text{down}}} + \frac{S_{\text{up}}(\mathcal{A})}{B_{\text{up}}} \\
 T_{\text{comp}}(\mathcal{A}) &= \max_j T_{\text{client}}^j(\mathcal{A}) + T_{\text{server}}(\mathcal{A}), & T_{\text{client}}^j(\mathcal{A}) &= R_{\text{comp}} T_{\text{sim}}^j(\mathcal{A}) + C_{\text{comp}}
 \end{aligned}$$

Since  $T_{\text{server}}(\mathcal{A})$  is relatively smaller than  $T_{\text{client}}^j(\mathcal{A})$ , we ignore  $T_{\text{server}}(\mathcal{A})$  in our experiments. We get client download size  $S_{\text{down}}(\mathcal{A})$  and upload size  $S_{\text{up}}(\mathcal{A})$  from the number of neural network parameters.  $\max_j T_{\text{sim}}^j(\mathcal{A})$  is the computation time in our simulation.

$$B_{\text{down}} \sim 0.75\text{MB/secs}, B_{\text{up}} \sim 0.25\text{B/secs}, R_{\text{comp}} \sim 7, C_{\text{comp}} \sim 10\text{secs}$$

[35] estimate each value of the above parameters from a real world cross-device FL system. The upload bandwidth  $B_{\text{up}}$  is generally smaller than download bandwidth  $B_{\text{down}}$ . We define human time as the parallel time estimated by this new time metric.

### 6.1.1 Training Models

For MNIST, we use a  $l_2$ -regularized logistic regression model for the strongly convex case and a multilayer perceptron (MLP) with two hidden layers for the non-convex case. For CIFAR-10, we use a Convolutional Neural Network (CNN). Here, we note that the number of parameters in a neural network model is directly related to the number of communicated bits. We discuss more details as follows.

**MLP Model for MNIST.** We use a multilayer perceptron (MLP) with two hidden layers. Each hidden layer consists of 200 neurons with ReLU activations. Thus, we compute the total number of parameters in this MLP model as below.

$$\begin{aligned}
 (\# \text{ of MLP parameters}) &= (\# \text{ of input features}) \times (\# \text{ of neurons in the 1st layer}) \\
 &\quad + (\# \text{ of neurons in the 1st layer}) \times (\# \text{ of neurons in the 2nd layer}) \\
 &\quad + (\# \text{ of neurons in the 2nd layer}) \times (\# \text{ of MNIST classes}) \\
 &\quad + (\# \text{ of neurons in the 1st layer}) + (\# \text{ of neurons in the 2nd layer}) \\
 &\quad + (\# \text{ of MNIST classes}) \\
 &= 28 \times 28 \times 200 + 200 \times 200 + 200 \times 10 + 200 + 200 + 10 = 199210
 \end{aligned}$$

Finally, we derive  $S_{\text{up}}(\mathcal{A}) (= S_{\text{down}}(\mathcal{A}))$ , defined in section 6.1 (New time metric), by using the above fact. We use 32 bits floating-point if there is no quantization.

$$\begin{aligned}
 S_{\text{up}}(\mathcal{A}) &= (\# \text{ of device}) \times (\# \text{ of MLP parameters}) \times (\# \text{ of bits}) \\
 &= 18 \times 199210 \times 32 = 114744960
 \end{aligned}$$

The FedAvg algorithm follows the above calculation. If we use 8 bits quantization for FedPAQ, FedCOMGATE, and FedAQ, ( $\#$  of bits) in the above equation will respectively be 8, 8, and 16. Since FedAQ sends twice as many messages as others at every communication round, ( $\#$  of bits) for FedAQ is 16. Similarly, ( $\#$  of bits) for FedAC, which has no quantization, is 64.

**CNN Model for CIFAR-10.** We use a CNN model, which consists of two 2-dimensional convolutional layers, two max pooling layers, and two fully connected layers. The ReLU activations are used in this CNN model. Let's clarify (# of input channel, # of output channel, kernel size, stride) for convolutional layers. We respectively use (3, 20, 5, 1), (20, 50, 5, 1) for the 1st and 2nd convolutional layer. Let's denote each convolutional layer and fully connected layer as CONV1, CONV2, FC3, FC4. At first, the activation shape of input layer for CIFAR-10 is (32, 32, 3). Then, we get the activation shape after CONV1 and the number of parameters for CONV1.

$$\begin{aligned}
(\text{width of activation shape}) &= \frac{(\text{width of previous activation shape}) - \text{kernel size} + 1}{\text{stride}} \\
&= \frac{32 - 5 + 1}{1} = 28 \Rightarrow \text{activation shape} = (28, 28, 20) \\
(\# \text{ of CONV1 parameters}) &= \left( \text{kernel size} \times \text{kernel size} \right. \\
&\quad \times (\# \text{ of filters in the previous layer}) + 1 \Big) \\
&\quad \times (\# \text{ of filters in the current layer}) \\
&= (5 \times 5 \times 3 + 1) \times 20 = 1520
\end{aligned}$$

The activation shape becomes (14, 14, 20) after max pooling. There are no learnable parameters in pooling layers. We do similar calculation for CONV2.

$$\begin{aligned}
(\text{width of activation shape}) &= \frac{(\text{width of previous activation shape}) - \text{kernel size} + 1}{\text{stride}} \\
&= \frac{14 - 5 + 1}{1} = 10 \Rightarrow \text{activation shape} = (10, 10, 50) \\
(\# \text{ of CONV2 parameters}) &= \left( \text{kernel size} \times \text{kernel size} \times (\# \text{ of filters in the previous layer}) \right. \\
&\quad \left. + 1 \right) \times (\# \text{ of filters in the current layer}) \\
&= (5 \times 5 \times 20 + 1) \times 50 = 25050
\end{aligned}$$

The activation shape becomes (5, 5, 50) after second max pooling. Then, we calculate the number of parameters in FC3 and FC4 similar to the MLP case.

$$\begin{aligned}
(\# \text{ of FC3 parameters}) &= (5 \times 5 \times 50) \times 512 + 512 = 640512 \\
(\# \text{ of FC4 parameters}) &= 512 \times 10 + 10 = 5130
\end{aligned}$$

Thus, the total number of parameters in this CNN model is

$$\begin{aligned}
(\# \text{ of CNN parameters}) &= (\# \text{ of CONV1 parameters}) + (\# \text{ of CONV2 parameters}) \\
&\quad + (\# \text{ of FC3 parameters}) + (\# \text{ of FC4 parameters}) \\
&= 1520 + 25050 + 640512 + 5130 = 672212
\end{aligned}$$

Finally, we derive  $S_{\text{up}}(\mathcal{A}) (= S_{\text{down}}(\mathcal{A}))$  in this case.

$$\begin{aligned}
S_{\text{up}}(\mathcal{A}) &= (\# \text{ of device}) \times (\# \text{ of CNN parameters}) \times (\# \text{ of bits}) \\
&= 8 \times 672212 \times 32 = 172086272
\end{aligned}$$

We can do the similar discussion in the MLP case when it comes to applying this to quantization-based federated optimization algorithms.

## 6.2 Experimental Results

In our experiments on both MNIST and CIFAR-10, we verify how the global training loss and test accuracy of five algorithms change with respect to communication rounds, the number of bits communicated between one client and the server during the uplink, and human time. We provide both qualitative analysis and quantitative results for plots.

### 6.2.1 Qualitative Analysis

**Strongly Convex Case.** In this experiment, we compare FedAQ under the condition set (34) and set (35) with FedAvg, FedPAQ, FedCOMGATE, and FedAC-I. We denote each FedAQ as FedAQ-I and FedAQ-II. As we observe the theoretical benefits of FedAQ over other methods in section 5, FedAQ-I outperforms all other quantization-based federated optimization algorithms and FedAC-I in all plots (See each first row of Figure 1, 2). However, although FedAQ-II shows the fast convergence speed, the training process is unstable. Thus, we only use FedAQ-I for further non-convex experiments. FedAC and FedAQ in non-convex experiments indicate FedAC-I and FedAQ-I.

**Non-Convex Case.** Each second row of Figure 1, 2, and Figure 3 clearly demonstrates that FedAQ with 4 bits quantization outperforms other algorithms in all plots. In terms of communication rounds, accelerated algorithms, FedAQ and FedAC, converge faster than other algorithms. We also observe that quantization does not lead to slower convergence, which means we can apply an efficient quantization scheme to make communication efficient FL systems without sacrificing convergence speed. The plots related to communicated bits are helpful to interpret how algorithms work well in situations with heavy communication. FedAQ with 8 bits quantization shows comparable performance relative to FedPAQ and FedCOMGATE with the help of acceleration, even though FedAQ sends more updates during every synchronization. When we use 4 bits quantization for FedAQ to make the number of communicated bits the same for all quantization-based algorithms during synchronization, FedAQ shows a much faster convergence speed with regard to the number of communicated bits. However, plots of communicated bits fail to reflect how algorithms converge in real estimated time for FL scenarios, which consists of both communication and computation. Thus, we further analyze algorithms with human time. We observe that FedAQ with 8 quantization bits performs slightly better than FedPAQ and FedCOMGATE for both MNIST and CIFAR-10. This occurs because while all quantization-based algorithms send the same number of communicated bits, the number of communication rounds for FedAQ is much smaller than others. Then, this also indicates that FedAQ takes less computation time than other methods while reaching the same accuracy.

### 6.2.2 Quantitative Results

We provide quantitative results to help readers understand plots better. To be specific, for all plots, we observe the number of communication rounds, the number of communicated bits, and the human time required to achieve a particular test accuracy by each federated optimization algorithm.

For the strongly convex experiment on MNIST (See the first row of Figure 2), the number of communication rounds required to achieve 90.28% test accuracy by FedAvg, FedPAQ(8bits), FedCOMGATE(8bits), FedAC-I, FedAQ-I(8bits), FedAQ-II(8bits) are respectively 217, 216, 260, 28, 26, 99. The number of communicated bits required to achieve the same accuracy are respectively  $5.4e7$ ,  $1.4e7$ ,  $1.6e7$ ,  $1.4e7$ ,  $3.3e6$ ,  $1.2e7$ . Lastly, the required human time are respectively 3220s, 2760s, 3336s, 484s, 344s, 1323s. In this experiment, FedAQ-I(8bits) requires the smallest number of communication rounds, the smallest number of communicated bits, and the shortest human time to achieve the same test accuracy. These experimental results support the validity of our theoretical analysis on strongly convex cases.

For the non-convex experiment on MNIST (See the second row of Figure 2), the number of communication rounds required to achieve 97.6% test accuracy by FedAvg, FedPAQ(8bits), FedCOMGATE(8bits), FedAC,



FedAQ(8bits), FedAQ(4bits) are respectively 23, 48, 38, 18, 18, 16. The number of communicated bits required to achieve the same accuracy are respectively  $1.5e8$ ,  $7.6e7$ ,  $6.1e7$ ,  $2.3e8$ ,  $5.7e7$ ,  $2.5e7$ . Finally, the required human time are respectively 2424s, 2311s, 1834s, 3327s, 1248s, 805s. Thus, we conclude that FedAQ(4bits) outperforms other algorithms, and even FedAQ(8bits) needs smaller number of communicated bits/less human time to achieve the goal accuracy than FedPAQ(8bits)/FedCOMGATE(8bits).

For the non-convex experiment on CIFAR-10 (See Figure 3), the number of communication rounds required to achieve 65.4% test accuracy by FedAvg, FedPAQ(8bits), FedCOMGATE(8bits), FedAC, FedAQ(8bits), FedAQ(4bits) are respectively 98, 89, 95, 49, 50, 48. The number of communicated bits required to achieve the same accuracy are respectively  $2.1e9$ ,  $4.8e8$ ,  $5.1e8$ ,  $2.1e9$ ,  $5.4e8$ ,  $2.6e8$ . Finally, the required human time are respectively 31798s, 11526s, 12240s, 28720s, 9902s, 6464s. As with the non-convex experiment on MNIST, FedAQ(4bits) outperforms other algorithms, and even FedAQ(8bits) requires less human time to achieve the same accuracy than FedPAQ(8bits)/FedCOMGATE(8bits).

**Remark 24:** Our current experimental setup only allows us to scale the number of clients up to the number of CPU cores in our machine. Since FedAQ achieves linear speed up in the number of workers with much fewer communication rounds than other quantization based methods, we expect FedAQ to outperform other methods by an even larger margin as we scale the number of workers.

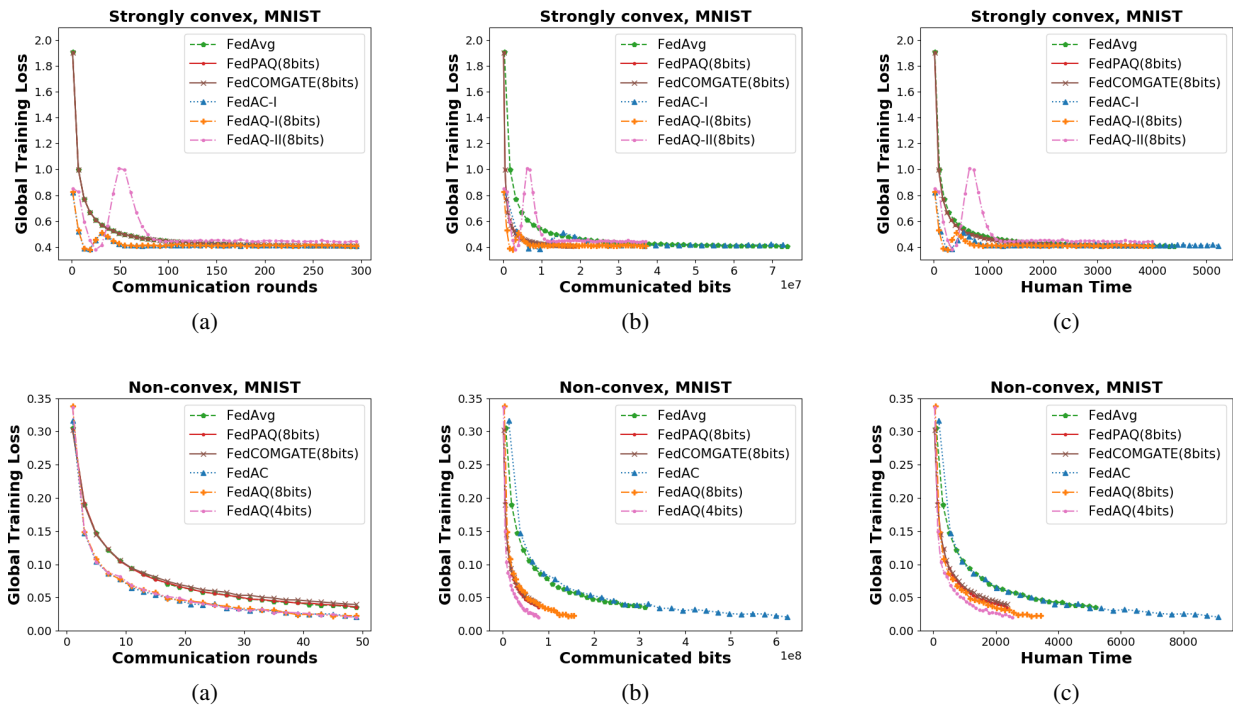


Figure 1: Comparing FedAQ with FedAvg, FedPAQ, FedCOMGATE, and FedAC on MNIST with Strongly Convex Settings (first row) and Non-Convex Settings (second row). We observe how the global training loss changes across communication rounds (first column), communicated bits (second column), and human time (third column). FedAQ-I(8bits) and FedAQ(4bits) respectively outperform other algorithms for strongly convex settings and non-convex settings. FedAQ(4bits) sends the same number of communicated bits as FedPAQ(8bits) and FedCOMGATE(8bits) in each communication round, which indicates a fair comparison (See Quantization bits in section 6.1).

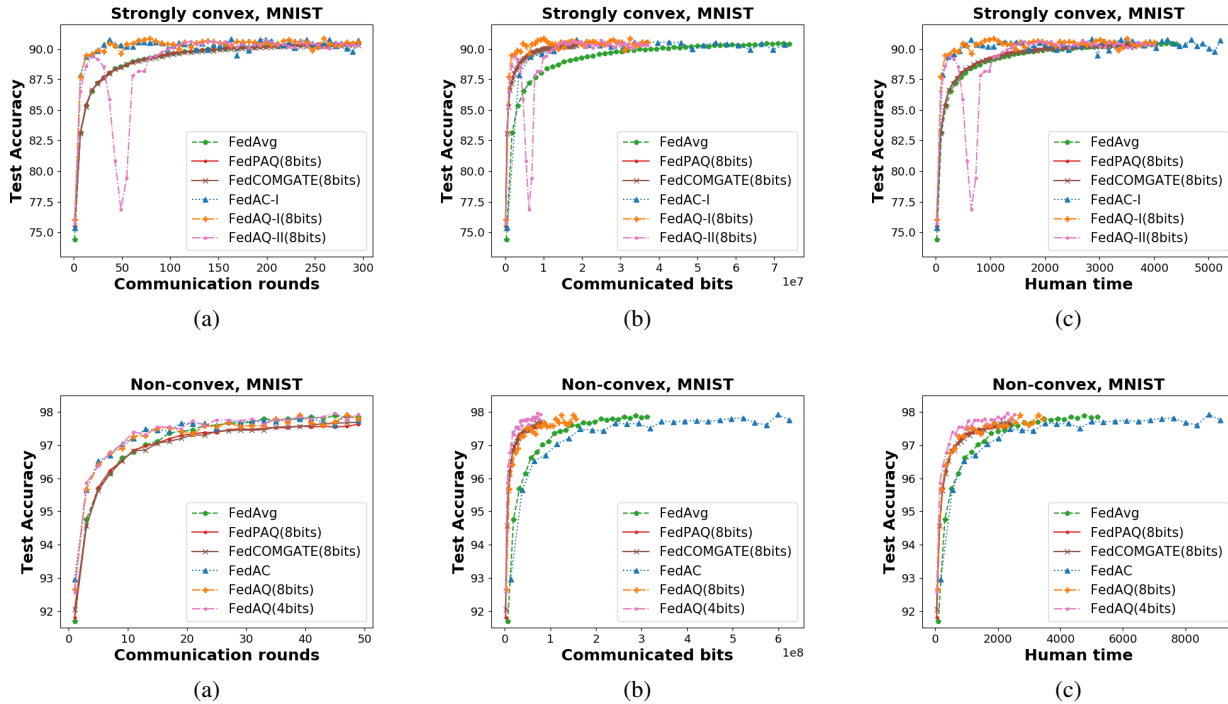


Figure 2: Comparing FedAQ with FedAvg, FedPAQ, FedCOMGATE, and FedAC on MNIST with Strongly Convex Settings (first row) and Non-Convex Settings (second row). We observe how the test accuracy changes across communication rounds (first column), communicated bits (second column), and human time (third column). FedAQ-I outperforms other algorithms in all plots for strongly convex settings. Moreover, FedAQ(4bits) outperforms other algorithms in all plots for non-convex settings.

## 7 Conclusion

To sum up, we propose a novel communication-efficient federated optimization algorithm, FedAQ, that successfully incorporates accelerated multiple local updates and quantization with solid theoretical guarantees in strongly-convex and homogeneous settings. In the future, further theoretical guarantees of FedAQ on convex and non-convex functions should be discussed. Also, the convergence analysis of FedAQ on heterogeneous settings can be an interesting topic. Even though Federated Learning systems provide some level of privacy to the clients as their explicit data is not shared with the servers, careful examination of FL systems including FedAQ is necessary to examine how much privacy do they actually provide as information is shared in form of the iterates.

## References

- [1] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.
- [2] N. Bansal and A. Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.
- [3] D. Basu, D. Data, C. Karakus, and S. Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations. *arXiv preprint arXiv:1906.02367*, 2019.

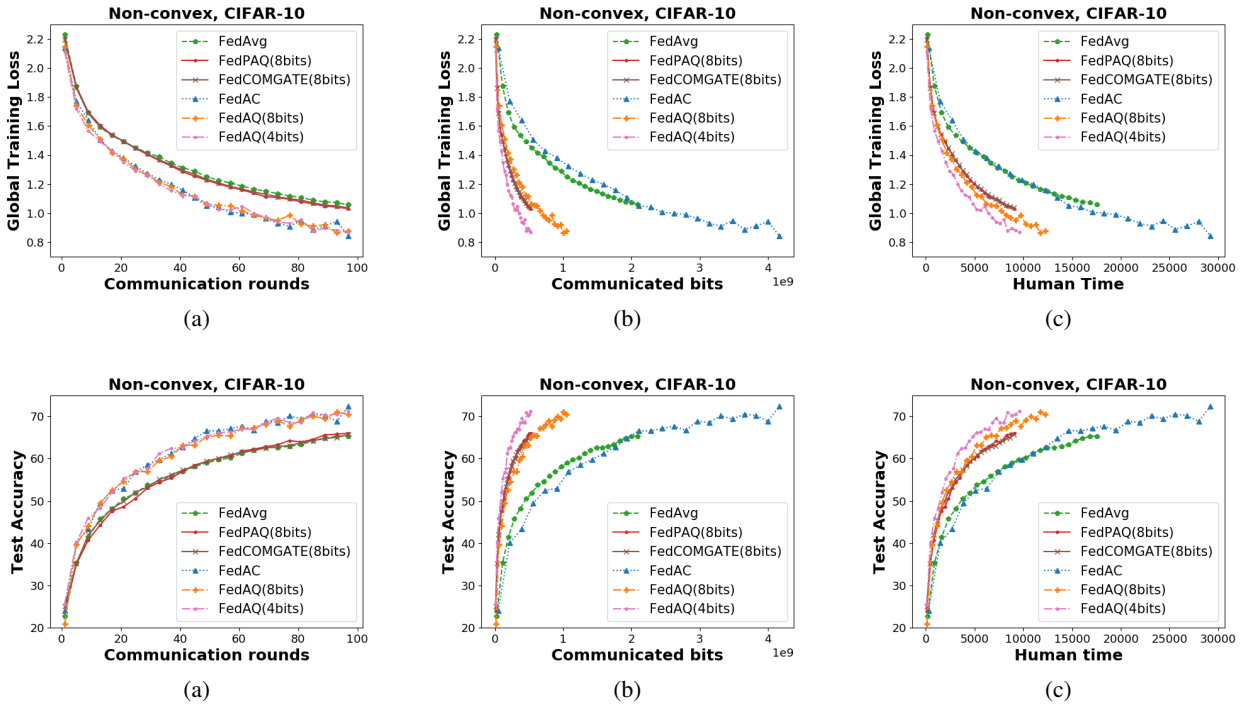


Figure 3: Comparing FedAQ with FedAvg, FedPAQ, FedCOMGATE, and FedAC on CIFAR-10. We observe how the global training loss and test accuracy change across communication rounds (first column), communicated bits (second column), and human time (third column). We use a CNN model for CIFAR-10. Similar to the MNIST experiment, FedAQ (4 bits) outperforms all other algorithms in every case.

- [4] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [6] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.
- [7] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, pages 11082–11094, 2019.
- [8] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe. Trading redundancy for communication: Speeding up distributed sgd for non-convex optimization. In *International Conference on Machine Learning*, pages 2545–2554. PMLR, 2019.
- [9] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2021.

- [10] F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. [arXiv preprint arXiv:1910.14425](#), 2019.
- [11] S. Horvath, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtárik. Natural compression for distributed deep learning. [arXiv preprint arXiv:1905.10988](#), 2019.
- [12] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. [arXiv preprint arXiv:1912.04977](#), 2019.
- [13] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. [arXiv preprint arXiv:2008.03606](#), 2020.
- [14] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In [International Conference on Machine Learning](#), pages 5132–5143. PMLR, 2020.
- [15] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In [International Conference on Artificial Intelligence and Statistics](#), pages 4519–4529. PMLR, 2020.
- [16] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. [arXiv preprint arXiv:1610.05492](#), 2016.
- [17] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Manuscript, 2009.
- [18] Y. LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [19] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. [IEEE Signal Processing Magazine](#), 37(3):50–60, 2020.
- [20] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. [arXiv preprint arXiv:1812.06127](#), 2018.
- [21] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. [arXiv preprint arXiv:1907.02189](#), 2019.
- [22] X. Li, B. Karimi, and P. Li. On distributed adaptive optimization with gradient compression. [arXiv preprint arXiv:2205.05632](#), 2022.
- [23] Z. Li, D. Kovalev, X. Qian, and P. Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. [arXiv preprint arXiv:2002.11364](#), 2020.
- [24] Z. Li and P. Richtárik. Canita: Faster rates for distributed convex optimization with communication compression. [arXiv preprint arXiv:2107.09461](#), 2021.
- [25] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don’t use large mini-batches, use local sgd. [arXiv preprint arXiv:1808.07217](#), 2018.
- [26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In [Artificial intelligence and statistics](#), pages 1273–1282. PMLR, 2017.
- [27] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In [International Conference on Artificial Intelligence and Statistics](#), pages 2021–2031. PMLR, 2020.

- [28] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora. Fetchsgd: Communication-efficient federated learning with sketching. In International Conference on Machine Learning, pages 8253–8265. PMLR, 2020.
- [29] N. Singh, D. Data, J. George, and S. Diggavi. Squarm-sgd: Communication-efficient momentum sgd for decentralized optimization. IEEE Journal on Selected Areas in Information Theory, 2021.
- [30] S. U. Stich. Local sgd converges fast and communicates little. arXiv preprint arXiv:1805.09767, 2018.
- [31] S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. arXiv preprint arXiv:1909.05350, 2019.
- [32] A. T. Suresh, X. Y. Felix, S. Kumar, and H. B. McMahan. Distributed mean estimation with limited communication. In International Conference on Machine Learning, pages 3329–3337. PMLR, 2017.
- [33] T. Vogels, S. P. Karimireddy, and M. Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. Advances In Neural Information Processing Systems 32 (Nips 2019), 32(CONF), 2019.
- [34] H. Wang, S. Sievert, Z. Charles, S. Liu, S. Wright, and D. Papailiopoulos. Atomo: Communication-efficient learning via atomic sparsification. arXiv preprint arXiv:1806.04090, 2018.
- [35] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. arXiv preprint arXiv:2107.06917, 2021.
- [36] J. Wang and G. Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. arXiv preprint arXiv:1808.07576, 2018.
- [37] J. Wang, Z. Xu, Z. Garrett, Z. Charles, L. Liu, and G. Joshi. Local adaptivity in federated learning: Convergence and consistency. arXiv preprint arXiv:2106.02305, 2021.
- [38] Y. Wang, L. Lin, and J. Chen. Communication-efficient adaptive federated learning. arXiv preprint arXiv:2205.02719, 2022.
- [39] J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. arXiv preprint arXiv:1710.09854, 2017.
- [40] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. McMahan, O. Shamir, and N. Srebro. Is local sgd better than minibatch sgd? In International Conference on Machine Learning, pages 10334–10343. PMLR, 2020.
- [41] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In International Conference on Machine Learning, pages 7184–7193. PMLR, 2019.
- [42] H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 5693–5700, 2019.
- [43] H. Yuan and T. Ma. Federated accelerated stochastic gradient descent. Advances in Neural Information Processing Systems, 33, 2020.