FADAS: Towards Federated Adaptive Asynchronous Optimization

Yujia Wang [1] **Shiqiang Wang** [2] **Songtao Lu** [2] **Jinghui Chen** [1]

Abstract

Federated learning (FL) has emerged as a widely adopted training paradigm for privacy-preserving machine learning. While the SGD-based FL algorithms have demonstrated considerable success in the past, there is a growing trend towards adopting adaptive federated optimization methods, particularly for training large-scale models. However, the conventional synchronous aggregation design poses a significant challenge to the practical deployment of those adaptive federated optimization methods, particularly in the presence of straggler clients. To fill this research gap, this paper introduces federated adaptive asynchronous optimization, named FADAS, a novel method that incorporates asynchronous updates into adaptive federated optimization with provable guarantees. To further enhance the efficiency and resilience of our proposed method in scenarios with significant asynchronous delays, we also extend FADAS with a delay-adaptive learning adjustment strategy. We rigorously establish the convergence rate of the proposed algorithms and empirical results demonstrate the superior performance of FADAS over other asynchronous FL baselines.

1. Introduction

In recent years, federated learning (FL) (McMahan et al., 2017) has drawn increasing attention as an efficient privacy-preserving distributed machine learning paradigm. An FL framework consists of a central server and numerous clients, where clients collaboratively train a global model without sharing their private data. FL entails each client conducting multiple local iterations, while the central server periodically aggregates these local updates into the global model. Following the original design of the FedAvg algorithm (McMahan et al., 2017), a large number of stochastic gradient descent (SGD)-based FL methods have emerged, aiming to improve the performance or efficiency of FedAvg (Karimireddy et al., 2020; Acar et al., 2021; Wang et al., 2020b).

In addition to the successes of SGD-based algorithms in enhancing the efficiency of FL, the adoption of adaptive optimization techniques is becoming increasingly prevalent in FL. Adaptive optimization techniques such as Adam (Kingma & Ba, 2015) and AdamW (Loshchilov & Hutter, 2017) have proven their advantages over SGD in effectively training or fine-tuning large-scale models like BERT (Devlin et al., 2018), ViT (Dosovitskiy et al., 2021), and Llama (Touvron et al., 2023). This progress has encouraged the incorporation of adaptive optimization into the FL settings, taking advantage of their ability to navigate update directions and dynamically adjust learning rates. For example, FedAdam (Reddi et al., 2021) and FedAMS (Wang et al., 2022b) employ global adaptive optimization after the server aggregates local model updates. Moreover, strategies such as FedLALR (Sun et al., 2023a), FedLADA (Sun et al., 2023b), and FAFED (Wu et al., 2023) replace SGD with the Adam optimizer for the local training phase, exemplifying the utility of local adaptive optimizations in FL.

However, existing methods in adaptive FL still rely on traditional synchronous aggregation approaches, where the server must wait for all participating clients to complete their local training before global updates. This reliance presents a significant challenge to the practical implementation of adaptive FL methodologies, as the server is required to wait until slower clients, which may have limited computation or communication capabilities. While asynchronous FL strategies such as FedBuff (Nguyen et al., 2022) and FedAsync (Xie et al., 2019) have been investigated to improve the scalability and to study the impact of client delays on the convergence of SGD-based FL algorithms, the specific implications of asynchronous delays on nonlinear adaptive gradient operations are not completely understood. This motivates us to explore the following question:

Can we develop an asynchronous method for adaptive federated optimization (with provable guarantees) that enhances training efficiency and is resilient to asynchronous delays?

[1]College of Information Sciences and Technology, Pennsylvania State University, State College, PA, USA [2]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Yujia Wang <yjw5427@psu.edu>, Shiqiang Wang <wangshiq@us.ibm.com>, Songtao Lu <songtao@ibm.com>, Jinghui Chen <jzc5917@psu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

In this paper, we propose FADAS, **F**ederated **AD**aptive **AS**ynchronous optimization, to address this challenge. FADAS introduces asynchronous updates within the adaptive federated optimization framework and integrates a delay-adaptive mechanism for adjusting the learning rate adaptively in response to burst delays. We summarize our contributions as follows:

- We propose FADAS, a novel adaptive federated optimization method that extends traditional adpative federated optimization *support asynchronous client updates*. We prove that FADAS achieves a convergence rate of $\mathcal{O}\big(\frac{1}{\sqrt{TM}} + \frac{\tau_{\max}\tau_{\text{avg}}}{T}\big)$ w.r.t. the number of global communication rounds T and the number of accumulated updates M, with bounded worst-case delay, denoted by $\tau_{\max}$, and the average of the maximum delay over all the rounds, denoted by τ_{avg}.
- To further reduce the dependency on the worst-case delay term $\tau_{\max}$ in the convergence rate, we extend FADAS with a *delay-adaptive learning rate adjustment strategy*. Our theoretical results demonstrate that the inclusion of a delay-adaptive learning rate effectively diminishes the dependency on $\tau_{\max}$ in the convergence rate.
- We conduct experiments across various asynchronous delay settings in both vision and language modeling tasks. Our results indicate that the proposed FADAS, whether or not including the delay-adaptive learning rate, outperforms other asynchronous FL baselines. In particular, the delay-adaptive FADAS demonstrates significant advantages in scenarios with large worst-case delays. Moreover, our experimental results on simulating the wall-clock training time underscores the efficiency of our proposed FADAS approach.

2. Related Work

Federated learning. FL, as introduced by McMahan et al. (2017), has become a pivotal framework for collaboratively training machine learning models on edge devices while keeping local data private. Following the initial FedAvg algorithm, several works studied the theoretical analysis and empirical performance of it (Lin et al., 2018; Stich, 2018; Li et al., 2019a; Karimireddy et al., 2020; Wang & Joshi, 2021; Yang et al., 2021), and a range of works aim to improve FedAvg from different perspectives, such as reducing the impact of data heterogeneity (Karimireddy et al., 2020; Acar et al., 2021; Wang et al., 2020b), saving the communication overhead (Reisizadeh et al., 2020; Jhunjhunwala et al., 2021), and adjusting the parameter aggregation procedure (Tan et al., 2022; Wang & Ji, 2023).

Adaptive FL optimizations and adaptive updates. Besides traditional SGD-based methods, there is a line of works focusing on adaptive updates in FL. A local adaptive FL method with momentum-based variance-reduced gradient was used in FAFED (Wu et al., 2023). Li et al. (2023) proposed a framework for local adaptive gradient methods in FedDA. FedLALR (Sun et al., 2023a) uses local adaptive optimization in FL with local historical gradients and periodically synchronized learning rates. FedLADA (Sun et al., 2023b) is an efficient local adaptive FL method with a locally amended technique. Jin et al. (2022) developed novel adaptive FL optimization methods from the perspective of dynamics of ordinary differential equations. Moreover, Reddi et al. (2021) introduced FedAdagrad, FedAdam and FedYogi, and Wang et al. (2022b) proposed FedAMS for global adaptive FL optimizations. Several works of global adaptive learning rate (Jhunjhunwala et al., 2023) and adaptation in aggregation weights (Tan et al., 2022; Wang & Ji, 2023) are also related to adaptive learning rate adjustment.

Asynchronous SGD and asynchronous FL. There have been extensive studies over the years about asynchronous optimization techniques, including asynchronous SGD and its various adaptations. For example, Hogwild (Niu et al., 2011) includes an applicable lock-free, coordinate-wise asynchronous method and has been widely used in multi-thread computation. A body of works focuses on the theoretical analysis and explorations of asynchronous SGD (Mania et al., 2017; Nguyen et al., 2018; Stich et al., 2021; Leblond et al., 2018; Glasgow & Wootters, 2022) and discusses the gradient delay in the convergence rate (Avdiukhin & Kasiviswanathan, 2021; Mishchenko et al., 2022; Koloskova et al., 2022; Wu et al., 2022). Within federated learning, innovative asynchronous aggregation algorithms like FedAsync (Xie et al., 2019) allow the server to update the global model once a client finishes local training, and FedBuff (Nguyen et al., 2022) introduces a buffered aggregation approach. There are also many works focusing on algorithms based on FedBuff with theoretical and/or empirical analysis (Toghani & Uribe, 2022; Ortega & Jafarkhani, 2023; Wang et al., 2023), and other aspects of asynchronous FL (Chen et al., 2020b; Yang et al., 2022; Bornstein et al., 2023). Although adaptive FL and asynchronous FL have achieved the success of training large machine learning models with desirable numerical performance, the exploration of asynchronous updates in the context of adaptive FL has not been well-studied yet. In this paper, we start with the asynchronous update framework in adaptive FL and further integrate delay-adaptive learning rate scheduling into it.

3. Preliminaries

Federated learning. A general FL framework considers a distributed optimization problem across N clients:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\boldsymbol{x};\xi_i)], \quad (1)$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the model parameter with d dimensions, $F_i(\boldsymbol{x})$ is the loss function corresponding to client i, $\mathcal{D}_i$ is the local data distribution on client i. The objective in Eq. (1) can be interpreted as setting $p_i = \frac{1}{N}$ for all clients in another commonly used objective function in FL, i.e., $f(\boldsymbol{x}) = \sum_{i=1}^N p_i \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_i(\boldsymbol{x}; \xi_i)]$ with $p_i \geq 0$ and $\sum_{i=1}^N p_i = 1$. FedAvg (McMahan et al., 2017) is a typical synchronous FL algorithm to solve Eq. 1, where in the t-th global round, each participating client i performs local SGD updates as follows:

$$\boldsymbol{x}_{t,k+1}^i = \boldsymbol{x}_{t,k}^i - \eta_l \nabla F_i(\boldsymbol{x}_{t,k}^i; \xi) \text{ and } \boldsymbol{x}_{t,0}^i = \boldsymbol{x}_t \quad (2)$$

where η_l is the learning rate. After several local steps (e.g., K steps of local training), the server performs a global averaging step after receiving all the updates from assigned clients in $\mathcal{S}_t$, i.e., $\boldsymbol{x}_{t+1} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \boldsymbol{x}_{t,K}^i$.

Adaptive optimization and its application to FL. Several adaptive optimizers have been proposed to improve the convergence of SGD, such as Adagrad (Duchi et al., 2011), RMSProp (Tieleman et al., 2012), Adam (Kingma & Ba, 2015) and its variant AMSGrad (Reddi et al., 2018). In general machine learning optimization, Adam effectively inherits the benefits of both momentum and RMSProp optimizers, leading to better empirical performance in practical applications.

Reddi et al. (2021) first introduced adaptive federated optimization, which applies the adaptive optimizers during the global aggregation steps in FL. FedAMSGrad (Tong et al., 2020) and FedAMS (Wang et al., 2022b) further adjust the effective global learning rate in adaptive FL. Specifically, FedAdam and FedAMS take the idea of viewing the difference of local updates $\boldsymbol{\Delta}_t^{\text{sync}} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \boldsymbol{\Delta}_t^{i,\text{sync}} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} (\boldsymbol{x}_{t,K}^i - \boldsymbol{x}_t)$ as a pseudo-gradient, and applies the Adam or AMSGrad optimizer when updating global model $\boldsymbol{x}_{t+1}$ using $\boldsymbol{\Delta}_t^{\text{sync}}$, i.e.,

$$\begin{aligned}
&\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{\Delta}_t^{\text{sync}}, \\
&\boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{\Delta}_t^{\text{sync}} \odot \boldsymbol{\Delta}_t^{\text{sync}}, \\
&\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t} + \epsilon} \text{ (FedAdam)}, \\
&\widehat{\boldsymbol{v}}_t = \max(\widehat{\boldsymbol{v}}_{t-1}, \boldsymbol{v}_t), \boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta \frac{\boldsymbol{m}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} \text{ (FedAMS)},
\end{aligned}$$

where $\odot$ denotes the element-wise product for two vectors, and for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\sqrt{\boldsymbol{x}}, \boldsymbol{x}/\boldsymbol{y}, \max(\boldsymbol{x}, \boldsymbol{y})$ denote the element-wise square root, division, and maximum operation of the vectors.

Asynchronous updates in FL. In asynchronous FL, clients train the model asynchronously and update it to the server once it finishes several steps of local training. FedBuff (Nguyen et al., 2022) has improved the global update steps with the concept of buffer based on the initial FedAsync baseline (Xie et al., 2019). In FedBuff, it requires the framework maintain a given number (referred to as the concurrency M_c) of clients that are actively local training. At the t-th global round, after the client i finishes local training, it sends its local update $\boldsymbol{\Delta}_t^i = \boldsymbol{x}_{t-\tau,K}^i - \boldsymbol{x}_{t-\tau}$ to the server, where $t - \tau$ is the global round where client i starts local training and $0 \leq \tau \leq t$. The server simultaneously accumulates the model update $\boldsymbol{\Delta}_t^i$ to the global update direction $\boldsymbol{\Delta}_t \leftarrow \boldsymbol{\Delta}_t + \boldsymbol{\Delta}_t^i$, and sends the latest global model to a randomly selected client who is idle. When the number of accumulated updates reaches the given buffer size of M, the server updates the global model with the averaging $\boldsymbol{\Delta}_t/M$. Meanwhile, clients who have not finished their local training will continue their training based on the previously received global model, and are not affected by the global model updates on the server. During the training, the framework always maintains a fixed number (M_c) of clients who are conducting local training. This is achieved by having the server randomly sample an idle client for training each time a client completes its local training and sends its update to the server.

Discussion about synchronous and asynchronous methods. Synchronous FL typically offers consistency and stability, i.e., all client updates are based on the same global model, and this consistency may lead to a more stable and predictable learning process. However, when there exist one or a few clients that are much slower than the majority of clients, which often happens in large-scale systems, synchronous FL can be inefficient since every client needs to wait for the slowest client before progressing with the next round of training. Asynchronous FL is more efficient when clients have system heterogeneity such as diverse computational capabilities or communication bandwidth. In FL, if the delay among clients is relatively uniform, synchronous FL tends to be more stable and efficient. Overall, the choice between synchronous and asynchronous FL hinges on specific needs and system characteristics. Synchronous FL is ideal in homogeneous systems, while asynchronous FL is advantageous in heterogeneous systems with potential straggler clients.

4. Proposed Method: FADAS

Although adaptive FL methods achieve promising convergence and generalization performance theoretically and empirically, the existing adaptive FL methods are restricted to synchronous settings, as the server needs to wait for all the assigned clients to finish their local updates for aggregation and then update the global model. However, those synchronous adaptive FL algorithms are susceptible to the presence of stragglers, where slower clients with insufficient computation or communication speed impede the progress of the global update.

To improve the efficiency and resiliency of adaptive FL in the presence of stragglers, we introduce FADAS, a **F**ederated **AD**aptive **AS**ynchronous optimization method. Similar to FedAdam and FedAMS, the proposed FADAS algorithm takes the model update difference from clients as a pseudo-gradient and it updates the global model following an Adam-like update scheme. Algorithm 1 summarizes the details. FADAS keeps the local asynchronous training scheme as FedBuff and maintains the concept of concurrency and buffer size for flexible control of the number of active clients and the frequency of global model update. In FADAS, after the server aggregates to obtain model update difference $\boldsymbol{\Delta}_t$, it finds an adaptive update direction, whose components are computed based on the AMSGrad optimizer (Reddi et al., 2018) as follows:

$$\begin{cases} \boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{\Delta}_t, \\ \boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{\Delta}_t \odot \boldsymbol{\Delta}_t, \\ \widehat{\boldsymbol{v}}_t = \max(\widehat{\boldsymbol{v}}_{t-1}, \boldsymbol{v}_t). \end{cases} \tag{3}$$

In general, FADAS enables clients to conduct local training in their own pace, and the server aggregates the asynchronous updates for global adaptive updates. It improves the training efficiency and scalability of over synchronous adaptive FL while inheriting the advantage of adaptive optimizer of reducing oscillations and stabilizing the optimization process.

Although FADAS applies asynchronous local training for adaptive FL, the global adaptive optimizer adjusts the global update direction only based on local updates but without considering the impact of asynchronous delay. Intuitively, a large asynchronous delay from a client means that this model update is made based on an outdated global model. This may lead to a negative effect on the convergence, and later we also verify this intuition in the theoretical analysis. This inspires us to apply a delay-adaptive learning rate adjustment to improve the resiliency of FADAS to stragglers with large delays. Specifically, we let the server track the delay for every received model update and adopt a delay-adaptive learning rate. We highlight the delay-adaptive steps in Algorithm 1 and those steps are executed with almost no extra overhead.

Delay tracking. In general, the server manages the delay record for each client through straightforward timestamping. For example, the server records the global update round t' when it broadcasts the current global model $\boldsymbol{x}_{t'}$ to client i, the client conducts local training with $\boldsymbol{x}_{t'}$. When the server receives the first $\boldsymbol{\Delta}_t^i$ from client i at round $t \geq t'$, the gradient delay for $\boldsymbol{\Delta}_t^i$, which is $\tau_t^i = t - t'$, is updated and recorded on the server.

Delay-adaptive learning rate. Assume that for each global update round t, clients in the set $\mathcal{M}_t$ ($|\mathcal{M}_t| = M$) send updates to the server. The received model updates at global round t have a maximum delay $\tau_t^{\max}$ defined as $\tau_t^{\max} := \max\{\tau_t^i, i \in \mathcal{M}_t\}$. Suppose we set up a delay threshold τ_c, we can define a delay-adaptive learning rate as:

$$\eta_t = \begin{cases} \eta & \text{if } \tau_t^{\max} \leq \tau_c, \\ \min\left\{\eta, \frac{1}{\tau_t^{\max}}\right\} & \text{if } \tau_t^{\max} > \tau_c. \end{cases} \tag{4}$$

Intuitively, this design implies that we need to turn the learning rates down for the model update $\boldsymbol{\Delta}_t$ with larger current-step delays. Specifically, if the current-step maximum delay $\tau_t^{\max}$ is larger than a given threshold τ_c, we scale down the learning rates for this step in proportional to $1/\tau_t^{\max}$ (also capped by a constant learning rate η) to avoid that the high-latency update worsens the convergence.

Comparison with FedAsync (Xie et al., 2019). FedAsync (Xie et al., 2019) also studies delay-adaptive weighted averaging during global model updates. In FedAsync, after the server receives a local model $\boldsymbol{x}_{\text{new}}$, it updates $\boldsymbol{x}_t$ based on $\boldsymbol{x}_t = (1-\alpha_t)\boldsymbol{x}_{t-1} + \alpha_t \boldsymbol{x}_{\text{new}}$, and FedAsync includes a hinge strategy of α_t which is similar to our delay-adaptive strategy in Eq. (4). However, unlike FedAsync, where the server updates the global model immediately upon receiving a new update from a client, FADAS updates the global model less frequently. In FADAS, the server accumulates

Algorithm 1 FADAS (with delay adaptation)

Input: local learning rate η_l, global learning rate η, adaptive optimization parameters $\beta_1, \beta_2, \epsilon$, server concurrency M_c, buffer size M, delay threshold τ_c ;

Initialize model $\boldsymbol{x}_1$, initialize $\boldsymbol{\Delta}_1 = \boldsymbol{0}$, $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{v}_0 = \boldsymbol{0}$, $m = 0$ and sample a set of $\mathcal{M}_0$ with size M_c active clients to run **local SGD updates**.
repeat
  if receive client update **then**
    Server accumulates update from client i: $\boldsymbol{\Delta}_t \leftarrow \boldsymbol{\Delta}_t + \boldsymbol{\Delta}_t^i$ and set $m \leftarrow m+1$
    Sample another client j from available clients
    Send the current model $\boldsymbol{x}_t$ to client j, and run **local SGD updates** on client j
  end if
  if $m = M$ **then**
    $\boldsymbol{\Delta}_t \leftarrow \frac{\boldsymbol{\Delta}_t}{M}$
    Update $\boldsymbol{m}_t, \boldsymbol{v}_t, \widehat{\boldsymbol{v}}_t$ by (3)
    if delay-adaptive **then**
      Set η_t to be delay-adaptive based on Eq. (4)
    else
      $\eta_t = \eta$
    end if
    Update global model $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \frac{\boldsymbol{m}_t}{\sqrt{\widehat{\boldsymbol{v}}_t + \epsilon}}$
    Set $m \leftarrow 0$, $\boldsymbol{\Delta}_{t+1} \leftarrow \boldsymbol{0}$, $t \leftarrow t+1$
  end if
until convergence

M local updates before a global update. Moreover, the convergence analysis in FedAsync did not consider their delay adaptation procedure, while we provide a convergence analysis incorporating the effect of delay adaptation in the next section.

5. Theoretical Analysis

In this section, we delve into the theoretical analysis of our proposed FADAS algorithm. We first introduce some common assumptions required for the analysis. Subsequently, we present the analysis in two parts: one focusing on FADAS without delay adaptation, as discussed in Section 5.1, and the other on the delay-adaptive FADAS in Section 5.2.

Assumption 5.1 (Smoothness)**.** Each objective function on the i-th worker $F_i(\boldsymbol{x})$ is L-smooth, i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Assumption 5.2 (Bounded Variance)**.** Each stochastic gradient is unbiased and has a bounded local variance, i.e., for all $\boldsymbol{x}, i \in [N]$, we have $\mathbb{E}\big[\|\nabla F_i(\boldsymbol{x};\xi) - \nabla F_i(\boldsymbol{x})\|^2\big] \leq \sigma^2$, and the loss function on each worker has a global variance bound, $\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma_g^2$.

Assumption 5.1 and 5.2 are standard assumptions in federated non-convex optimization literature (Li et al., 2019b; Yang et al., 2021; Reddi et al., 2021; Wang et al., 2022b; Wang & Ji, 2023). The global variance upper bound of σ_g^2 in Assumption 5.2 measures the data heterogeneity across clients, and a global variance of $\sigma_g^2 = 0$ indicates a uniform data distribution across clients.

Assumption 5.3 (Bounded Gradient)**.** Each loss function on the i-th worker $F_i(\boldsymbol{x})$ has G-bounded stochastic gradient on ℓ_2 norm, i.e., for all ξ, we have $\|\nabla F_i(\boldsymbol{x};\xi)\| \leq G$.

Assumption 5.3 is necessary for adaptive gradient algorithms for both general (Kingma & Ba, 2015; Chen et al., 2020a), distributed (Wang et al., 2022a) and federated adaptive optimization (Reddi et al., 2021; Wang et al., 2022b; Sun et al., 2023b). This is because the effective global learning rate for adaptive gradient methods is $\frac{\eta}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}$, and we need a lower bound for $\big\|\frac{\eta}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\big\|$ to guarantee that the effective learning rate does not vanish to zero.

Assumption 5.4 (Bounded Delay of Gradient Computation)**.** Let τ_t^i represent the delay for global round t and client i which is applied in Algorithm 1. The delay τ_t^i is the difference between the current global round t and the global round at which client i started to compute the gradient. We assume that the maximum gradient delay (worst-case delay) is bounded, i.e., $\tau_{\max} = \max_{t\in[T], i\in[N]}\{\tau_t^i\} < \infty$.

Assumption 5.4 is common in analyzing asynchronous and anarchic FL algorithms which incorporate the gradient delays into their algorithm design (Koloskova et al., 2022; Yang et al., 2021; Nguyen et al., 2022; Toghani & Uribe, 2022; Wang et al., 2023).

Assumption 5.5 (Uniform Arrivals of Gradient Computation)**.** Let the set $\mathcal{M}_t$ (with size M) include clients that transmit their local updates to the server in global round t. We assume that the clients' update arrivals are uniformly distributed, i.e., from a theoretical perspective, the M clients in $\mathcal{M}_t$ are randomly sampled without replacement from all clients $[N]$ according to a uniform distribution[1].

Assumption 5.5 is also discussed in Anarchic FL (Yang et al., 2022), which has been utilized to analyze the AFA-CD algorithm proposed therein.

5.1. Convergence Rate of FADAS

For expository convenience, in the following, we provide the theoretical convergence analysis of FADAS under the case of $\beta_1 = 0$. The theoretical analysis and the proof for the general case of $0 \leq \beta_1 < 1$ are provided in Appendix A. We define the average of the maximum delay over time as $\tau_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T}\tau_t^{\max} = \frac{1}{T}\sum_{t=1}^{T}\max_{i\in[N]}\{\tau_t^i\}$ which is useful in our analysis.

Theorem 5.6. *Under Assumptions 5.1–5.5, let T represent the total number of global rounds, K be the number of local SGD training steps and M be the number of the accumulated updates (buffer size) in each round. If the learning rate η and η_l satisfies $\eta\eta_l \leq \min\left\{\frac{\epsilon^2 M(N-1)}{180 C_G N(N-M)\tau_{\max} KL}, \frac{\sqrt{\epsilon^3 M(N-1)}}{12\sqrt{C_G N(M-1)\tau_{\max}^3} KL}\right\}$, $\eta_l \leq \frac{\sqrt{\epsilon}}{\sqrt{360 C_G \tau_{\max}} KL}$, then the global iterates $\{\boldsymbol{x}_t\}_{t=1}^T$ of Algorithm 1 satisfy*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&\leq \frac{4C_G}{\eta\eta_l KT}\mathcal{F} + \frac{20 C_G \eta_l^2 K L^2(\sigma^2 + 6K\sigma_g^2)}{\epsilon} \\
&\quad + \left[\frac{8C_G\eta^2\eta_l^2 K L^2 \tau_{\text{avg}}\tau_{\max}}{M\epsilon^3} + \frac{12 C_G \eta\eta_l L}{M\epsilon^2}\right] \\
&\quad \cdot \left\{\sigma^2 + \frac{N-M}{N-1}[15\eta_l^2 K^2 L^2(\sigma^2 + 6K\sigma_g^2) + 3K\sigma_g^2]\right\},
\end{aligned} \tag{5}$$

where $\mathcal{F} = f(\boldsymbol{x}_1) - f_*$, $f_* = \min_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$ and $C_G = \eta_l K G + \epsilon$.

Corollary 5.7. *If we choose the global learning rate $\eta = \Theta(\sqrt{M})$ and $\eta_l = \Theta\Big(\frac{\sqrt{\mathcal{F}}}{\sqrt{TK(\sigma^2 + K\sigma_g^2)}}\Big)$ in Theorem 5.6, then for sufficiently large T, the global iterates $\{\boldsymbol{x}_t\}_{t=1}^T$*

[1]This assumption is only used for theoretical analysis. Our experiments that show the advantage of FADAS empirically do not rely on this assumption.

of Algorithm 1 satisfy

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \mathcal{O}\Big(\frac{\sqrt{\mathcal{F}}\sigma}{\sqrt{TKM}} + \frac{\sqrt{\mathcal{F}}\sigma_g}{\sqrt{TM}} + \frac{\mathcal{F}}{T} + \frac{\mathcal{F}G}{T\sqrt{M}} + \frac{\mathcal{F}\tau_{\max}\tau_{\text{avg}}}{T}\Big), \tag{6}$$

Remark 5.8. Corollary 5.7 suggests that given sufficiently large T and relatively small worst-case delay $\tau_{\max}$, the proposed FADAS (without delay-adaptive learning rate) achieves a convergence rate of $\mathcal{O}\big(\frac{1}{\sqrt{TM}}\big)$ w.r.t. T and M.

Comparison to asynchronous FL methods. Compared with the analysis for FedBuff in Nguyen et al. (2022) and Toghani & Uribe (2022), our analysis for FADAS obtains a relaxed dependency on the worst-case gradient delay $\tau_{\max}$, and FADAS achieves a slightly better rate on non-dominant term than $\mathcal{O}\big(\frac{1}{\sqrt{T}} + \frac{\tau_{\max}^2}{T}\big)$ obtained in Toghani & Uribe (2022). Moreover, Wang et al. (2023) also studied the convergence for FedBuff with relaxed requirements for $\tau_{\max}$, and our FADAS achieves a similar convergence of $\mathcal{O}\big(\frac{1}{\sqrt{TM}} + \frac{\tau_{\max}\tau_{\text{avg}}}{T}\big)$ as in Wang et al. (2023). It is worthwhile to mention that recently CA^{2}FL (Wang et al., 2023) improves the convergence of asynchronous FL under heterogeneous data distributions, while the improvement is obtained by using the cached variable on the server for global update calibration.

Note that when $\tau_{\max}$ in Eq. (6) is large, particularly in cases where $\tau_{\max} \geq \frac{\sqrt{T}}{\sqrt{M}}$, then $\frac{\tau_{\max}\tau_{\text{avg}}}{T}$ becomes the dominant term in the convergence rate. This implies that a large worst-case delay $\tau_{\max}$ may lead to a worse convergence rate. In the next subsection, we demonstrate that the delay-adaptive learning rate strategy can relieve this problem and enhance FADAS with better resilience to large worst-case delays.

5.2. Convergence Rate of Delay-adaptive FADAS

In the following, we provide the convergence analysis for delay-adaptive FADAS with $\beta_1 = 0$. To get started, we first define the median of the maximum delay over all communication rounds $[T]$:

$$\tau_{\text{median}} = \text{median}\{\tau_1^{\max}, \tau_2^{\max}, ..., \tau_T^{\max}\}. \tag{7}$$

The definition of τ_{median} implies that the number of global update rounds that have a maximum delay greater than τ_{median} is less than half of the total number of global updates T. With this definition, we present the following theorem characterizing the convergence rate of delay-adaptive FADAS.

Theorem 5.9. *Under Assumptions 5.1–5.5, let T be the total number of global rounds, K be the number of local SGD training steps and M be the number of the buffer size in each round. If the learning rate η and η_l satisfies $\eta\eta_l \leq \min\Big\{\frac{\epsilon^2 M(N-1)}{60C_G N(N-M)\tau_{\max}KL}, \frac{\sqrt{\epsilon^3 M(N-1)}}{12\sqrt{C_G N(M-1)\tau_{\max}^3}KL}\Big\}$, $\eta_l \leq \frac{\sqrt{\epsilon}}{\sqrt{360C_G\tau_{\max}}KL}$ and $\eta \leq \frac{\sqrt{M}}{\tau_c}$, then the global iterates $\{\boldsymbol{x}_t\}_{t=1}^T$ of Algorithm 1 satisfy*

$$\begin{aligned}
&\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&\leq \frac{4C_G}{\eta\eta_l KT}\mathcal{F} + \frac{20C_G\eta_l^2KL^2(\sigma^2+6K\sigma_g^2)}{\epsilon} \\
&+ \frac{8C_G\eta^3\eta_l^2KL^2T\tau_{\text{avg}}}{M\epsilon^3\sum_{t=1}^T\eta_t}\sigma^2 + \frac{8C_G\eta^2\eta_l^2KL^2T\tau_{\text{avg}}}{\sqrt{M}\epsilon^3\sum_{t=1}^T\eta_t} \\
&\cdot \frac{N-M}{N-1}[15\eta_l^2K^2L^2(\sigma^2+6K\sigma_g^2)+3K\sigma_g^2] + \frac{4C_G\eta\eta_l L}{M\epsilon^2} \\
&\cdot \Big\{\sigma^2 + \frac{N-M}{N-1}[15\eta_l^2K^2L^2(\sigma^2+6K\sigma_g^2)+3K\sigma_g^2]\Big\},
\end{aligned} \tag{8}$$

where $\mathcal{F} = f(\boldsymbol{x}_1) - f_*$, $f_* = \min_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$ and $C_G = \eta_l KG + \epsilon$.

Corollary 5.10. *If we pick $\tau_c = \tau_{\text{median}}$, the global learning rate $\eta = \Theta(\sqrt{M}/\tau_c)$ and $\eta_l = \Theta\big(\frac{\tau_c\sqrt{\mathcal{F}}}{\sqrt{TK(\sigma^2+K\sigma_g^2)}}\big)$, then for sufficiently large T, the global iterates $\{\boldsymbol{x}_t\}_{t=1}^T$ of Algorithm 1 satisfy*

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \mathcal{O}\Big(\frac{\sqrt{\mathcal{F}}\sigma}{\sqrt{TKM}} + \frac{\sqrt{\mathcal{F}}\sigma_g}{\sqrt{TM}} + \frac{\mathcal{F}G\tau_c}{T\sqrt{M}} + \frac{\mathcal{F}\tau_{\text{avg}}}{T} + \frac{\mathcal{F}(\tau_c^2+\tau_c\tau_{\text{avg}})}{T}\Big). \tag{9}$$

Remark 5.11. Corollary 5.10 suggests that with sufficiently large T, delay-adaptive FADAS also achieves a convergence rate of $\mathcal{O}\big(\frac{1}{\sqrt{TM}}\big)$ w.r.t. T and M.

Remark 5.12. Compared to the convergence rate in Corollary 5.7, the convergence rate in Corollary 5.10 does not rely on the (possibly large) worst-case delay $\tau_{\max}$. In cases where $\tau_c = \tau_{\text{median}} \approx \tau_{\text{avg}} \ll \tau_{\max}$, Corollary 5.10 relaxes the requirement from $\tau_{\max}$ to τ_{median} for achieving the desired convergence rate. Since τ_{median} describes the median of $\tau_t^{\max} = \max_{i\in[N]}\{\tau_t^i\}$ in each round t, the convergence rate in Corollary 5.10 is less sensitive to stragglers who may cause a large worst-case delay in the system.

6. Experiments

We explore the performance of our proposed FADAS algorithm through experiments on vision and language tasks, using the CIFAR-10/100 (Krizhevsky et al., 2009) datasets with ResNet-18 model (He et al., 2016) for vision tasks, and applying the pre-trained BERT base model (Devlin et al., 2018) for fine-tuning several datasets from

the GLUE benchmark dataset (Wang et al., 2018) for language tasks. We compare our proposed FADAS algorithm against asynchronous FL baselines, such as FedBuff (without differential privacy) (Nguyen et al., 2022) and FedAsync (Xie et al., 2019), a synchronous SGD-based FL baseline FedAvg (McMahan et al., 2017), and a synchronous adaptive FL baseline FedAMS (Wang et al., 2022b). We summarize some crucial implementation details in the following, and we leave some additional results and experiment details to Appendix D. Our code can be found at `https://github.com/yujiaw98/FADAS`.

Overview of vision tasks' implementation. We set up a total of 100 clients for the *mild delay* scenario, in which the concurrency $M_c = 20$ and the buffer size $M = 10$ by default. We also set up a total of 50 clients for the *large worst-case delay* scenario, with $M_c = 25$ and $M = 5$ correspondingly. For both settings, we partition the data on clients based on the Dirichlet distribution following Wang et al. (2020a;b), and the parameter α used in Dirichlet sampling determines the degree of data heterogeneity. We apply two levels of data heterogeneity with $\alpha = 0.1$ and $\alpha = 0.3$. Each client conducts two local epochs of training, and the mini-batch size is 50 for each client. The local optimizer for all methods is SGD with weight decay 10^{-4}, and we grid search the global and local learning rates individually for each method.

Overview of language tasks' implementation. Considering the total number of data samples in the language classification datasets, we set up a total of 10 clients, partition the data on clients based on the labels, and we apply a heterogeneity level of $\alpha = 0.6$. Each client conducts one local epoch and the mini-batch size is 32 for each client. The local optimizer for all methods is SGD with weight decay 10^{-4}, and we grid search the global and local learning rates individually for each method. We set the concurrency $M_c = 5$ and buffer size $M = 3$ by default. We employ the widely-used low-rank adaptation method, LoRA (Hu et al., 2021), as a parameter-efficient fine-tuning strategy for our language classification tasks. This involves freezing the original pre-trained weight matrix $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$ and fine-tuning $\Delta\boldsymbol{W}$ through low-rank decomposition, where $\boldsymbol{W} = \boldsymbol{W}_0 + \alpha_{\text{LoRA}}\Delta\boldsymbol{W} = \boldsymbol{W}_0 + \alpha_{\text{LoRA}}\boldsymbol{B}\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{R}^{d\times r}$, and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$, and we adopt $r = 1$ and $\alpha_{\text{LoRA}} = 8$ in our experiments.

Overview of delay simulation. In our experiments, we simulate the asynchronous environment as follows. Initially, we partition clients into three categories, including `Small`, `Medium`, and `Large` delay, at the start of training and tag them with a label reflective of their delay magnitude. This partitioning was executed via a Dirichlet sampling process controlled by the parameter γ. A smaller γ value corresponds to a higher proportion of clients experiencing large delays. Unless otherwise specified in subsequent experiments, we set $\gamma = 1$. To mimic actual wall-clock running times within each delay category, we apply uniform sampling at each round for each client. We adopt the following uniform distributions to simulate wall-clock running time for both the *large worst-case delay* and *mild delay* settings as shown in Table 1.

6.1. Results on Vision Tasks

Large worst-case delay. Under this setting, we simulate the wall-clock running time by letting a small proportion of clients have more significant delays than other clients. Tables 2 and 3 show the overall performance of training the ResNet-18 model on CIFAR-10 and CIFAR-100, respectively. The results show that FADAS, especially with a delay-adaptive learning rate, offers significant advantages in terms of test accuracy. Compared to FedAsync and FedBuff, both FADAS methods achieve higher accuracy, and FADAS with delay-adaptive learning rates is shown to be more stable during the learning process with lower standard derivation. In these experiments, we conduct a total of $T = 500$ global communication rounds, and the maximum delay $\tau_{\max} = 127$, which even more than a quarter of the total number of global communication rounds. Notably, as seen in Tables 2 and 3, FedAsync shows severely fluctuating in test accuracy, suggesting that it may be less reliable in situations with large worst-case delays.

Table 1. Overview for wall-clock delay simulation (in units of 10 seconds).

Delay	`Small`	`Medium`	`Large`
Large worst-case	$U(1,2)$	$U(3,5)$	$U(50,80)$
Mild	$U(1,2)$	$U(3,5)$	$U(5,8)$

Mild delay. Under this setting, we simulate the wall-clock running time for clients by assuming that all clients can finish their local training within a comparable duration (see Table 1). Tables 4 and 5 show the overall performance of training the ResNet-18 model on CIFAR-10 and CIFAR-100 under mild delay. The results highlight that both FADAS and its delay-adaptive variant achieve superior test accuracy than FedAsync and FedBuff.

6.2. Results on Language Tasks

The performance for fine-tuning the BERT base model on three GLUE benchmark datasets, RTE, MRPC, and SST-2, under mild delay conditions are shown in Table 6, which illustrates that FADAS and its delay-adaptive counterpart consistently outperform the results of FedAsync and FedBuff across the three datasets. FedAsync achieves good performance in SST-2 but is less satisfactory in RTE and MRPC, and FedBuff presents an overall lower accu-

Table 2. The test accuracy on training ResNet-18 model on CIFAR-10 dataset with two data heterogeneity levels in a *large worst-case* delay scenario for 500 communication rounds. We report the average accuracy and standard derivation over the last 5 rounds, and we abbreviate delay-adaptive FADAS to FADAS_{da} in this and subsequent tables.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	50.92 ± 5.03	75.3 ± 6.18
FedBuff	38.68 ± 8.16	51.32 ± 4.43
FADAS	72.0 ± 0.94	73.27 ± 1.37
FADAS_{da}	$\mathbf{73.96} \pm 3.54$	$\mathbf{79.68} \pm 2.14$

Table 3. The test accuracy on training ResNet-18 model on CIFAR-100 dataset with two data heterogeneity levels in a *large worst-case* delay scenario for 500 communication rounds.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	46.51 ± 4.76	38.55 ± 7.36
FedBuff	13.04 ± 5.5	18.63 ± 5.13
FADAS	47.84 ± 0.59	53.64 ± 0.52
FADAS_{da}	$\mathbf{50.31} \pm 1.0$	$\mathbf{57.18} \pm 0.31$

racy with larger standard derivation compared with FADAS. The delay-adaptive FADAS shows parity with the standard FADAS algorithm under mild delays. Moreover, FADAS achieves significant accuracy improvements on RTE and MRPC datasets against the SGD-based asynchronous FL baselines, further demonstrating the intuition of developing the FADAS method.

Running time speedup. Table 7 demonstrates the efficiency of FADAS and its delay-adaptive variant by comparing their performance with two synchronous FL methods in reaching the target validation accuracy across different dataset. Notably, FADAS consistently outperforms FedAvg and FedAMS in terms of wall-clock running time, requiring significantly fewer time units to reach the desired accuracy levels. In vision classification tasks such as CIFAR-10 and CIFAR-100, the standard FADAS shows a significant reduction in training time, achieving $8\times$ speedup than FedAvg and more than $2.5\times$ speedup than FedAMS. The delay-adaptive FADAS shows similar results as the standard version. For language classification tasks, FADAS also improves the training time compared with FedAMS and FedAvg. These results highlight the scalability and efficiency of FADAS, especially when considering the computational constraints in practical FL environments.

6.3. Ablation studies

Sensitivity of delay adaptive learning rates. Figure 1 (a) exhibits the ablation study for different delay threshold τ_c for the delay-adaptive FADAS under the scenario of *large worst-case delays*. Following Eq. (4), τ_c provides a threshold so that we reduce the learning rate if there exists a client with extremely large delay. The experiment compares the accuracy of three thresholds $\tau_c = 1, 4, 8, 10$, and $\tau_c = 4$ shows very similar test accuracy as $\tau_c = 10$. The result in Figure 1 (a) shows that using $\tau_c = 8$ obtains a slightly better result than using $\tau_c = 1$, $\tau_c = 4$, and $\tau_c = 10$. It is interesting that in this *large worst-case delays* setting, we observe the average of the maximum delay $\tau_{\text{avg}} = 10.89$, the median of the maximum delay $\tau_{\text{median}} = 6.0$, and maximum delay during training is $\tau_{\max} = 127$, which shows $\tau_{\text{median}} \approx \tau_{\text{avg}} \ll \tau_{\max}$, confirming the practicality of our analysis as discussed in Remark 5.12. Together with the theoretical and experimental results, we find that the optimal choice of τ_c may depend on the actual delay during training.

Table 4. The test accuracy on training ResNet-18 model on CIFAR-10 dataset with two data heterogeneity levels under *mild delay* scenario.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	42.48 ± 4.93	71.76 ± 3.85
FedBuff	72.15 ± 2.71	79.82 ± 3.25
FADAS	77.68 ± 2.32	82.93 ± 0.81
FADAS_{da}	$\mathbf{78.93} \pm 0.83$	$\mathbf{83.91} \pm 0.54$

Table 5. The test accuracy on training ResNet-18 model on CIFAR-100 dataset with two data heterogeneity levels under *mild delay* scenario.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	45.26 ± 7.04	53.41 ± 8.94
FedBuff	53.70 ± 1.13	56.26 ± 1.64
FADAS	$\mathbf{57.37} \pm 0.47$	$\mathbf{61.22} \pm 0.31$
FADAS_{da}	57.21 ± 0.45	60.34 ± 0.42

Ablation for concurrency M_c and buffer size M. Figure 1 (b) presents the test accuracy of both the standard and delay-adaptive FADAS for different concurrency levels M_c, given the same buffer size $M = 5$. The delay-adaptive FADAS achieves higher accuracy than FADAS when concurrency $M_c = 15$ and $M_c = 25$, and the delay-adaptive version has worse accuracy at larger concurrency $M_c = 35$.

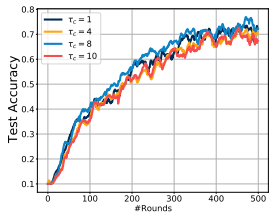
Table 6. The test accuracy on parameter-efficient fine-tuning BERT base model on three datasets from GLUE benchmark with heterogeneous data partitioned and *mild delay*.

Method	RTE Acc. & std.	MRPC Acc. & std.	SST-2 Acc. & std.
FedAsync	49.46 ± 2.66	69.71 ± 1.02	90.02 ± 0.79
FedBuff	61.61 ± 4.90	76.80 ± 6.05	78.37 ± 4.86
FADAS	64.26 ± 2.30	$\mathbf{83.33} \pm 1.20$	$\mathbf{90.76} \pm 0.26$
FADAS_{da}	$\mathbf{65.10} \pm 2.40$	83.09 ± 1.71	90.05 ± 1.80

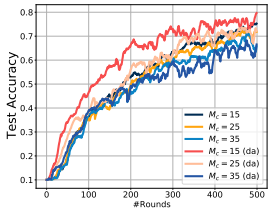
Table 7. Training/fine-tuning time simulation (in units of 10 seconds) to reach target test accuracy on the server under *mild delay* scenarios. For each dataset, the concurrency M_c is fixed for fair comparison.

	Acc.	FedAvg	FedAMS	FADAS	FADAS_{da}
CIFAR-10	75%	2257.7	648.7	**228.0**	237.5
CIFAR-100	50%	1806.3	546.9	**209.8**	**209.8**
RTE	63%	921.9	412.4	**376.2**	436.9
MRPC	80%	1018.1	424.0	**368.3**	370.1
SST-2	90%	-	495.2	73.8	**57.2**

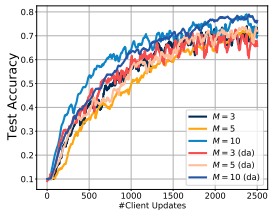
Figure 1 (c) presents an ablation study on buffer size for our proposed FADAS algorithm. It compares the performance of buffer sizes from $M = 3, 5, 10$ with their delay-adaptive counterparts over total client updates, i.e., the number of times the server receives updates from clients. It shows that with the same number of client trips, increasing the buffer size M tends to achieve higher accuracy. This is also due to the design of the concurrency-buffer size framework, as increasing the buffer size moves closer to traditional synchronous FL algorithms, i.e., clients are more likely to get up-to-date with the server. We also provide the comparison w.r.t. global communication round in Figure 1 (d). Figure 1 (d) shows that as the buffer size M increases, i.e., the number of clients contributing to one step of global update increases, the test accuracy also increases.



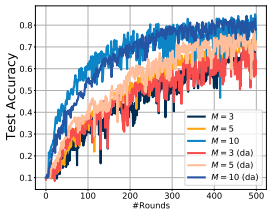
(a) Ablation on τ_c



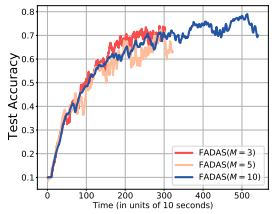
(b) Ablation on concurrency



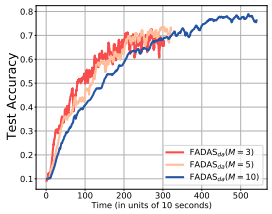
(c) Ablation on buffer size



(d) Ablation on buffer size



(e) Run time for FADAS



(f) Run time for FADAS_{da}

Figure 1. Several ablation studies based on training ResNet-18 model on CIFAR-10 data under *large worst-case* delay setting.

Moreover, we simulate the running time (similar to the setting for Table 7) for different buffer sizes M to investigate the time efficiency for adopting different buffer sizes. Figure 1 (e) and (f) show the run time for FADAS and delay-adaptive FADAS. They reveal that a smaller buffer size ($M = 3$) may have less training time to achieve a target accuracy, e.g., 70%. These results demonstrate that using smaller buffer sizes may yield higher accuracy in the early stage of training. In conjunction with the results shown in Figure 1 (c) and (d), we think there is a trade-off between the time of reaching some initial target accuracy (that is slightly lower than the final accuracy) and the final accuracy with regard to the buffer size. A larger buffer size M may yield improved final accuracy at convergence, but it also means that the server needs to wait for slower clients and there are less frequent updates of the global model, so the training speed at initial rounds can be slower.

7. Conclusion

In this paper, we propose FADAS, a novel asynchronous FL method that addresses the challenges of asynchronous updates in adaptive federated optimization. Based on the standard FADAS, we further integrate delay-adaptive learning rates to enhance the resiliency to stragglers with large delays. We theoretically establish the convergence rate for both standard and delay-adaptive FADAS under nonconvex stochastic settings. Our theoretical analysis indicates that the delay-adaptive algorithm substantially reduces the impact of severe worst-case delays on the convergence rate. Empirical evaluations across multiple tasks affirm that FADAS outperforms existing asynchronous FL methods and offers improved training efficiency compared to synchronous adaptive FL methods.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work is partially supported by the National Science Foundation under Grant No. 2348541. The views

and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

Impact Statement

This paper will make long-lasting contributions to the field of asynchronous federated learning and adaptive optimization. The focus of this work is on the technical advancement and optimization development of federated learning algorithms, and while there are numerous potential societal impacts of machine learning at large, this research does not necessitate a specific discussion on the societal consequences.

References

Acar, D. A. E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., and Saligrama, V. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Avdiukhin, D. and Kasiviswanathan, S. Federated learning under arbitrary communication patterns. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 425–435, 2021.

Bornstein, M., Rabbani, T., Wang, E. Z., Bedi, A., and Huang, F. SWIFT: Rapid decentralized federated learning via wait-free model communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jh1nCir1R3d.

Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020a.

Chen, T., Jin, X., Sun, Y., and Yin, W. Vafl: a method of vertical asynchronous federated learning. *arXiv preprint arXiv:2007.06081*, 2020b.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Glasgow, M. R. and Wootters, M. Asynchronous distributed optimization with stochastic delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 9247–9279. PMLR, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jhunjhunwala, D., Gadhikar, A., Joshi, G., and Eldar, Y. C. Adaptive quantization of model updates for communication-efficient federated learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3110–3114. IEEE, 2021.

Jhunjhunwala, D., Wang, S., and Joshi, G. FedExP: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023.

Jin, J., Ren, J., Zhou, Y., Lyu, L., Liu, J., and Dou, D. Accelerated federated learning with decoupled adaptive optimization. In *International Conference on Machine Learning*, pp. 10298–10322. PMLR, 2022.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Koloskova, A., Stich, S. U., and Jaggi, M. Sharper convergence guarantees for asynchronous SGD for distributed and federated learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4_oCZgBIVI.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Leblond, R., Pedregosa, F., and Lacoste-Julien, S. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *Journal of Machine Learning Research*, 19(81):1–68, 2018.

Li, J., Huang, F., and Huang, H. Fedda: Faster framework of local adaptive gradient methods via restarted dual averaging. *arXiv preprint arXiv:2302.06103*, 2023.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019a.

Li, X., Yang, W., Wang, S., and Zhang, Z. Communication-efficient local decentralized SGD methods. *arXiv preprint arXiv:1910.09126*, 2019b.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mishchenko, K., Bach, F., Even, M., and Woodworth, B. E. Asynchronous SGD beats minibatch SGD under arbitrary delays. *Advances in Neural Information Processing Systems*, 35:420–433, 2022.

Nguyen, J., Malik, K., Zhan, H., Yousefpour, A., Rabbat, M., Malek, M., and Huba, D. Federated learning with buffered asynchronous aggregation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3581–3607. PMLR, 2022.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and hogwild! Convergence without the bounded gradients assumption. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3750–3758. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/nguyen18c.html.

Niu, F., Recht, B., Re, C., and Wright, S. J. Hogwild! a lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 693–701, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Ortega, T. and Jafarkhani, H. Asynchronous federated learning with bidirectional quantized communications and buffered aggregation. *arXiv preprint arXiv:2308.00263*, 2023.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

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

Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031. PMLR, 2020.

Stich, S., Mohtashami, A., and Jaggi, M. Critical parameters for scalable distributed learning with large batches and asynchronous updates. In *International Conference on Artificial Intelligence and Statistics*, pp. 4042–4050. PMLR, 2021.

Stich, S. U. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Sun, H., Shen, L., Chen, S., Sun, J., Li, J., Sun, G., and Tao, D. Fedlalr: Client-specific adaptive learning rates achieve linear speedup for non-iid data. *arXiv preprint arXiv:2309.09719*, 2023a.

Sun, Y., Shen, L., Sun, H., Ding, L., and Tao, D. Efficient federated learning via local adaptive amended optimizer with linear speedup. *arXiv preprint arXiv:2308.00522*, 2023b.

Tan, L., Zhang, X., Zhou, Y., Che, X., Hu, M., Chen, X., and Wu, D. Adafed: Optimizing participation-aware federated learning with adaptive aggregation weights. *IEEE Transactions on Network Science and Engineering*, 9(4): 2708–2720, 2022.

Tieleman, T., Hinton, G., et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Toghani, M. T. and Uribe, C. A. Unbounded gradients in federated learning with buffered asynchronous aggregation. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8. IEEE, 2022.

Tong, Q., Liang, G., and Bi, J. Effective federated adaptive gradient methods with non-iid decentralized data. *arXiv preprint arXiv:2009.06557*, 2020.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BkluqlSFDS.

Wang, J. and Joshi, G. Cooperative sgd: A unified framework for the design and analysis of local-update SGD algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758, 2021.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020b.

Wang, S. and Ji, M. A lightweight method for tackling unknown participation probabilities in federated averaging. *arXiv preprint arXiv:2306.03401*, 2023.

Wang, Y., Lin, L., and Chen, J. Communication-compressed adaptive gradient method for distributed nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 6292–6320. PMLR, 2022a.

Wang, Y., Lin, L., and Chen, J. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22802–22838. PMLR, 2022b.

Wang, Y., Cao, Y., Wu, J., Chen, R., and Chen, J. Tackling the data heterogeneity in asynchronous federated learning with cached update calibration. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Wu, X., Magnusson, S., Feyzmahdavian, H. R., and Johansson, M. Delay-adaptive step-sizes for asynchronous learning. In *International Conference on Machine Learning*, pp. 24093–24113. PMLR, 2022.

Wu, X., Huang, F., Hu, Z., and Huang, H. Faster adaptive federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10379–10387, 2023.

Xie, C., Koyejo, S., and Gupta, I. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

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

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

A. Convergence analysis for adaptive asynchronous FL

Proof of Theorem 5.6. Here we directly start with general $\beta_1 \geq 0$ cases. Following several previous works studied centralized and federated adaptive methods (Chen et al., 2018; Wang et al., 2022b), we adopt an auxiliary Lyapunov sequence $\boldsymbol{z}_t$, and assume $\boldsymbol{x}_0 = \boldsymbol{x}_1$, then for each $t \geq 1$, we have

$$\boldsymbol{z}_t = \boldsymbol{x}_t + \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) = \frac{1}{1-\beta_1}\boldsymbol{x}_t - \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_{t-1}. \tag{10}$$

For the difference between $\boldsymbol{z}_{t+1}$ and $\boldsymbol{z}_t$, we have

$$\begin{aligned}
\boldsymbol{z}_{t+1} - \boldsymbol{z}_t &= \frac{1}{1-\beta_1}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) - \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) \\
&= \frac{1}{1-\beta_1} \cdot \eta \frac{\boldsymbol{m}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{\beta_1}{1-\beta_1} \cdot \eta \frac{\boldsymbol{m}_{t-1}}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}} + \epsilon} \\
&= \frac{1}{1-\beta_1} \cdot \eta \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}[\beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{\Delta}_t] - \frac{\beta_1}{1-\beta_1} \cdot \eta \frac{\boldsymbol{m}_{t-1}}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}} + \epsilon} \\
&= \eta \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{\beta_1}{1-\beta_1} \cdot \left(\frac{\eta}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{\eta}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}} + \epsilon} \right) \boldsymbol{m}_{t-1},
\end{aligned} \tag{11}$$

where $\boldsymbol{\Delta}_t = -\frac{\eta_l}{M}\sum_{i\in\mathcal{M}_t}\sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\tau^i_t,k} = -\frac{\eta_l}{M}\sum_{i\in\mathcal{M}_t}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k};\xi)$ and $\mathcal{M}_t$ be the set that include client send the local updates to the server at global round t.

From Assumption 5.1, f is L-smooth, taking the total expectation over all previous round, $0, 1, ..., t-1$ on the auxiliary sequence $\boldsymbol{z}_t$,

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{t+1}) - f(\boldsymbol{z}_t)] \\
&= \mathbb{E}[f(\boldsymbol{z}_{t+1})] - f(\boldsymbol{z}_t)] \\
&\leq \mathbb{E}[\langle \nabla f(\boldsymbol{z}_t), \boldsymbol{z}_{t+1} - \boldsymbol{z}_t \rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_t\|^2] \\
&= \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \eta \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} \right\rangle\right]}_{I_1} \underbrace{-\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{z}_t), \frac{\beta_1}{1-\beta_1} \cdot \eta \left(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}} + \epsilon} \right) \boldsymbol{m}_{t-1} \right\rangle\right]}_{I_2} \\
&\quad + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}\left[\left\| \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{\beta_1}{1-\beta_1}\left(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}} + \epsilon} \right) \boldsymbol{m}_{t-1} \right\|^2\right]}_{I_3} \\
&\quad + \underbrace{\mathbb{E}\left[\left\langle (\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)), \eta \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} \right\rangle\right]}_{I_4}.
\end{aligned} \tag{12}$$

Bounding I_1 Denote a sequence $\bar{\boldsymbol{\Delta}}_t = -\frac{\eta_l}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\tau^i_t,k} = -\frac{\eta_l}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k};\xi)$, where $\xi \sim \mathcal{D}_i$. For I_1, there is

$$\begin{aligned}
I_1 &= \eta\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} \right\rangle\right] \\
&= \eta\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \frac{\bar{\boldsymbol{\Delta}}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon} \right\rangle\right] \\
&= \eta\mathbb{E}\left[\left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, \bar{\boldsymbol{\Delta}}_t + \eta_l K \nabla f(\boldsymbol{x}_t) - \eta_l K \nabla f(\boldsymbol{x}_t) \right\rangle\right] \\
&= -\eta\eta_l K\mathbb{E}\left[\left\| \frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}} \right\|^2\right] + \eta\mathbb{E}\left[\left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, \bar{\boldsymbol{\Delta}}_t + \eta_l K \nabla f(\boldsymbol{x}_t) \right\rangle\right]
\end{aligned}$$

$$= -\eta\eta_l K\mathbb{E}\bigg[\bigg\|\frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\bigg\|^2\bigg] + \eta\mathbb{E}\bigg[\bigg\langle\frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}, -\frac{\eta_l}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k};\xi_i) + \frac{\eta_l K}{N}\sum_{i\in[N]}\nabla F_i(\boldsymbol{x}_t)\bigg\rangle\bigg], \quad (13)$$

where the second equality holds due to the characteristic of uniform arrivals (see Assumption 5.4), thus $\mathbb{E}(\boldsymbol{\Delta}_t) = \bar{\boldsymbol{\Delta}}_t$. The last inequality holds by the definition of $\bar{\boldsymbol{\Delta}}_t$ and the fact of the objective function $f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N F_i(\boldsymbol{x})$. By the fact of $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}[\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a}-\boldsymbol{b}\|^2]$, for second term in (13), we have

$$\begin{aligned}
&\eta\mathbb{E}\bigg[\bigg\langle\frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}, -\frac{\eta_l}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1}\boldsymbol{g}^i_{t-\tau^i_t,k} + \frac{\eta_l K}{N}\sum_{i\in[N]}\nabla F_i(\boldsymbol{x}_t)\bigg\rangle\bigg] \\
&= \eta\mathbb{E}\bigg[\bigg\langle\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\nabla f(\boldsymbol{x}_t), -\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\frac{1}{NK}\sum_{i\in[N]}\sum_{k=0}^{K-1}(\boldsymbol{g}^i_{t-\tau^i_t,k} - \nabla F_i(\boldsymbol{x}_t))\bigg\rangle\bigg] \\
&= \eta\mathbb{E}\bigg[\bigg\langle\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\nabla f(\boldsymbol{x}_t), -\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\frac{1}{NK}\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\bigg\rangle\bigg] \\
&= \frac{\eta\eta_l K}{2}\mathbb{E}\bigg[\bigg\|\frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\bigg\|^2\bigg] + \frac{\eta\eta_l}{2N^2K}\mathbb{E}\bigg[\bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\bigg\|^2\bigg] \\
&\quad - \frac{\eta\eta_l}{2N^2K}\mathbb{E}\bigg[\bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg], \qquad (14)
\end{aligned}$$

where the second equality holds by $\mathbb{E}[\boldsymbol{g}^i_{t-\tau^i_t,k}] = \mathbb{E}[\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})]$. Then for the second term in Eq. (14) , we have

$$\begin{aligned}
&\frac{\eta\eta_l}{2N^2K}\mathbb{E}\bigg[\bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\bigg\|^2\bigg] \\
&\le \frac{\eta\eta_l}{2N^2K\epsilon}\mathbb{E}\bigg[\bigg\|\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\bigg\|^2\bigg] \\
&\le \frac{\eta\eta_l}{2N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2] \\
&\le \frac{\eta\eta_l}{N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\bigg[\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\tau^i_t})\|^2] + \mathbb{E}[\|\nabla F_i(\boldsymbol{x}_{t-\tau^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2]\bigg] \\
&\le \frac{\eta\eta_l}{N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\bigg[L^2\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-\tau^i_t}\|^2] + L^2\mathbb{E}[\|\boldsymbol{x}_{t-\tau^i_t} - \boldsymbol{x}^i_{t-\tau^i_t,k}\|^2]\bigg], \qquad (15)
\end{aligned}$$

where the second inequality holds by $\forall \boldsymbol{a}_i, \|\sum_{i=1}^n \boldsymbol{a}_i\|^2 \le n\sum_{i=1}^n\|\boldsymbol{a}_i\|^2$, and the last inequality holds by Assumption 5.1. For the second term in Eq. (15), following by Lemma C.5, there is

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{t-\tau^i_t} - \boldsymbol{x}^i_{t-\tau^i_t,k}\|^2] &= \mathbb{E}\bigg[\bigg\|\sum_{m=0}^{k-1}\eta_l\boldsymbol{g}^i_{t-\tau^i_t,m}\bigg\|^2\bigg] \\
&\le 5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]. \qquad (16)
\end{aligned}$$

For the first term in Eq. (15), since by $\forall \boldsymbol{a}_i, \|\sum_{i=1}^n \boldsymbol{a}_i\|^2 \le n\sum_{i=1}^n\|\boldsymbol{a}_i\|^2$, there is

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-\tau^i_t}\|^2] = \mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}(\boldsymbol{x}_{s+1} - \boldsymbol{x}_s)\bigg\|^2\bigg] \le \tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2] \le \tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\mathbb{E}\bigg[\bigg\|\eta\frac{\boldsymbol{m}_s}{\sqrt{\widehat{\boldsymbol{v}}_s}+\epsilon}\bigg\|^2\bigg], \quad (17)$$

then by decomposing stochastic noise,

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-\tau_t^i}\|^2] \\
&\leq \frac{\eta^2 \tau_t^i}{\epsilon^2} \sum_{s=t-\tau_t^i}^{t-1} \mathbb{E}[\mathbb{E}_s \|\boldsymbol{m}_s\|^2]] \\
&= \frac{\eta^2 \tau_t^i}{\epsilon^2} \sum_{s=t-\tau_t^i}^{t-1} \mathbb{E}\Big[\mathbb{E}_s\Big[\Big\|(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{1}{M}\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\eta_l[\boldsymbol{g}^j_{u-\tau_u^j,k} - \nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k}) + \nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})]\Big\|^2\Big] \\
&\leq \frac{2\eta^2\tau_t^i}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{K\eta_l^2}{M}\sigma^2 + \frac{2\eta^2\tau_t^i}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{\eta_l^2}{M^2}\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\Big\|^2\Big] \\
&\leq \frac{2\eta^2(\tau_t^i)^2}{\epsilon^2}\frac{K\eta_l^2}{M}\sigma^2 + \frac{2\eta^2\tau_t^i}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{\eta_l^2}{M^2}\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\Big\|^2\Big],
\end{aligned} \tag{18}
$$

where the first inequality holds by decomposing the momentum $\boldsymbol{m}_s$, i.e., $\boldsymbol{m}_s = (1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\boldsymbol{\Delta}_u = (1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{1}{M}\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\eta_l\boldsymbol{g}^j_{u-\tau_u^j,k}$. The second inequality holds by $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$ and the fact of $\mathbb{E}[\cdot]] = \mathbb{E}[\cdot]$, and the third inequality holds by $(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u} \leq 1$.

Following Lemma C.2, $\frac{1}{C_G}\|\boldsymbol{x}\| \leq \left\|\frac{\boldsymbol{x}}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\| \leq \frac{1}{\epsilon}\|\boldsymbol{x}\|$ and $C_G = \eta_l K G + \epsilon$, plugging Eq. (14), Eq. (15) and Eq. (16) to (13), we have

$$
\begin{aligned}
\mathbb{E}[I_1] \leq &-\frac{\eta\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2KC_G}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big] \\
&+ \frac{\eta\eta_l K L^2}{\epsilon}\Big[5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\Big] + \frac{2\eta_l^3\eta^3K^2L^2}{M\epsilon^3}\sigma^2\frac{1}{N}\sum_{i=1}^{N}(\tau_t^i)^2 \\
&+ \frac{2\eta_l^3\eta^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\Big\|^2\Big].
\end{aligned} \tag{19}
$$

Bounding I_2

$$
\begin{aligned}
I_2 &= -\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{z}_t), \frac{\beta_1}{1-\beta_1}\cdot\eta\Big(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big)\boldsymbol{m}_{t-1}\Big\rangle\Big] \\
&= -\eta\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t), \frac{\beta_1}{1-\beta_1}\Big(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big)\boldsymbol{m}_{t-1}\Big\rangle\Big] \\
&\leq \eta\mathbb{E}\Big[\|\nabla f(\boldsymbol{x}_t)\|\Big\|\frac{\beta_1}{1-\beta_1}\Big(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big)\boldsymbol{m}_{t-1}\Big\|\Big] \\
&\quad + \eta^2 L\mathbb{E}\Big[\Big\|\frac{\beta_1}{1-\beta_1}\frac{\boldsymbol{m}_{t-1}}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big\|\cdot\Big\|\frac{\beta_1}{1-\beta_1}\Big(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big)\boldsymbol{m}_{t-1}\Big\|\Big] \\
&\leq \frac{\beta_1}{1-\beta_1}\eta_l\eta KG^2\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big\|_1\Big] + \frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta_l^2\eta^2K^2G^2L\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big\|_1\Big],
\end{aligned} \tag{20}
$$

where the first inequality holds by $\langle\boldsymbol{a},\boldsymbol{b}\rangle \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$ and L-smoothness of f, i.e., $\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)\| \leq L\|\boldsymbol{z}_t - \boldsymbol{x}_t\|$, and by the definition of $\boldsymbol{z}_t$, there is $\boldsymbol{z}_t - \boldsymbol{x}_t = \frac{\beta_1}{1-\beta_1}\frac{\boldsymbol{m}_{t-1}}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}$. The second inequality holds by Lemma C.2.

Bounding I_3

$$
I_3 = \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{\beta_1}{1-\beta_1}\Big(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\Big)\boldsymbol{m}_{t-1}\Big\|^2\Big]
$$

$$
\begin{aligned}
&\leq \eta^2 L \mathbb{E}\left[\left\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|^2\right] + \eta^2 L \mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\left(\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right)\boldsymbol{m}_{t-1}\right\|^2\right] \\
&\leq \eta^2 L \mathbb{E}\left[\left\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|^2\right] + \eta^2 L \frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2 K^2 G^2 \mathbb{E}\left[\left\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right\|^2\right] \\
&\leq \frac{\eta^2 L}{\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] + \eta^2 L \frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2 K^2 G^2 \mathbb{E}\left[\left\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right\|^2\right],
\end{aligned} \tag{21}
$$

where the first inequality follows by Cauchy-Schwarz inequality, i.e., $\forall \boldsymbol{a}_i, \|\sum_{i=1}^n \boldsymbol{a}_i\|^2 \leq n\sum_{i=1}^n \|\boldsymbol{a}_i\|^2$, and the second one holds by Lemma C.2.

Bounding I_4

$$
\begin{aligned}
I_4 &= \mathbb{E}\left[\left\langle (\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)), \eta\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\rangle\right] \\
&\leq \mathbb{E}\left[\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)\|\left\|\eta\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|\right] \\
&\leq L\mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{x}_t\|\left\|\eta\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|\right] \\
&\leq \frac{\eta^2 L}{2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\frac{\boldsymbol{m}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|^2\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\left\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\right\|^2\right] \\
&\leq \frac{\eta^2 L}{2\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|\boldsymbol{m}_t\|^2] + \frac{\eta^2 L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2],
\end{aligned} \tag{22}
$$

where the second inequality holds by Assumption 5.1 (the L-smoothness of f), and the third inequality holds by the definition of $\boldsymbol{z}_t$ and the inequality $\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$.

Merging pieces. Therefore, by merging pieces together, we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)] = \mathbb{E}[I_1 + I_2 + I_3 + I_4] \\
\leq& -\frac{\eta\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2KC_G}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\right\|^2\right] \\
&+ \frac{\eta\eta_l K L^2}{\epsilon}\left[5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^N\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\right] + \frac{2\eta^3\eta_l^3 K^2 L^2}{M\epsilon^3}\sigma^2\frac{1}{N}\sum_{i=1}^N(\tau_t^i)^2 \\
&+ \frac{2\eta^3\eta_l^3 K L^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\mathbb{E}\left[\left\|\sum_{j\in\mathcal{M}_u}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\right\|^2\right] \\
&+ \frac{\beta_1}{1-\beta_1}\eta\eta_l K G^2\mathbb{E}\left[\left\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right\|_1\right] + \frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta^2\eta_l^2 K^2 G^2 L\mathbb{E}\left[\left\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right\|_1\right] \\
&+ \frac{\eta^2 L}{\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] + \frac{\beta_1^2}{(1-\beta_1)^2}\eta^2\eta_l^2 K^2 G^2 L\mathbb{E}\left[\left\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}\right\|^2\right] \\
&+ \frac{\eta^2 L}{2\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|\boldsymbol{m}_t\|^2] + \frac{\eta^2 L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2].
\end{aligned} \tag{23}
$$

Denote a few sequences: $\boldsymbol{G}_t = \sum_{j\in\mathcal{M}_t}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_t^j,k})$ and $\boldsymbol{V}_t = \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1}}+\epsilon}$, then re-write and organize the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{t+1}) - f(\boldsymbol{z}_t)] \\
\leq& -\frac{\eta\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2KC_G}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\eta\eta_l KL^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^N\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\right]+\frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2\frac{1}{N}\sum_{i=1}^N(\tau_t^i)^2\\
&+\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\mathbb{E}[\|\boldsymbol{G}_u\|^2]\\
&+\frac{\beta_1}{1-\beta_1}\eta\eta_lKG^2\mathbb{E}[\|\boldsymbol{V}_t\|_1]+\frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta^2\eta_l^2K^2G^2L\mathbb{E}[\|\boldsymbol{V}_t\|_1]+\frac{\beta_1^2}{(1-\beta_1)^2}\eta^2\eta_l^2K^2G^2L\mathbb{E}[\|\boldsymbol{V}_t\|^2]\\
&+\frac{3\eta^2L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2]+\frac{\eta^2L}{2\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|\boldsymbol{m}_t\|^2].
\end{aligned}
\tag{24}
$$

Summing over $t=1$ to T, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{T+1})-f(\boldsymbol{z}_1)]\\
\le&-\frac{\eta\eta_lK}{2C_G}\sum_{t=1}^T\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]+\frac{\eta\eta_lKL^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2\eta_l^2\frac{1}{N}\sum_{t=1}^T\sum_{i=1}^N\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\right]\\
&+\underbrace{\frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T(\tau_t^i)^2}_{A_0}+\underbrace{\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\mathbb{E}[\|\boldsymbol{G}_u\|^2]}_{A_1}\\
&+\left(\frac{\beta_1}{1-\beta_1}\eta\eta_lKG^2+\frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta^2\eta_l^2K^2G^2L\right)\sum_{t=1}^T\mathbb{E}[\|\boldsymbol{V}_t\|_1]+\frac{\beta_1^2}{(1-\beta_1)^2}\eta^2\eta_l^2K^2G^2L\sum_{t=1}^T\mathbb{E}[\|\boldsymbol{V}_t\|^2]\\
&+\frac{3\eta^2L}{2\epsilon^2}\sum_{t=1}^T\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2]+\frac{\eta^2L}{2\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\sum_{t=1}^T\mathbb{E}[\|\boldsymbol{m}_t\|^2]\\
&-\frac{\eta_l}{2KC_G}\sum_{t=1}^T\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\right\|^2\right],
\end{aligned}
\tag{25}
$$

we have the following for term A_0,

$$
A_0=\frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T(\tau_t^i)^2\le\frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2\tau_{\text{avg}}\tau_{\max}T.
\tag{26}
$$

By Lemma C.6, we have the following for term A_1,

$$
\begin{aligned}
A_1=&\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\mathbb{E}[\|\boldsymbol{G}_u\|^2]\\
=&\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\Bigg\{\frac{3M(N-M)}{N-1}\\
&\cdot\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+(30K^4L^2\eta_l^2+K^2)\frac{1}{N}\sum_{j=1}^N\mathbb{E}[\|\nabla f(\boldsymbol{x}_{u-\tau_u^j})\|^2]+K^2\sigma_g^2\right]\\
&+\frac{M(M-1)}{N(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^N\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\right\|^2\right]\Bigg\}\\
=&\underbrace{\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^s\beta_1^{s-u}\left[\frac{3M(N-M)}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\right]}_{A_2}
\end{aligned}
$$

$$+ \underbrace{\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\left[\frac{3M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{u-\tau_u^j})\|^2]\right]}_{A_3}$$

$$+ \underbrace{\frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\left\{\frac{M(M-1)}{N(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{u-\tau_u^j,k})\right\|^2\right]\right\}}_{A_4}$$

$$= A_2 + A_3 + A_4. \tag{27}$$

For term A_2, then re-organizing it we have

$$\begin{aligned} A_2 &\leq \frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tau_t^i)^2\left[\frac{3M(N-M)}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\right] \\ &\leq \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\left[\frac{6M(N-M)}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\right]T\tau_{\text{avg}}\tau_{\max}. \end{aligned} \tag{28}$$

For term A_3, we have

$$\begin{aligned} A_3 &\leq \frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\left\{\frac{3M(N-M)}{N-1}\cdot(30K^4L^2\eta_l^2+K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{u-\tau_u^j})\|^2]\right\} \\ &\leq \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^2\sum_{t=1}^{T}\left\{\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^j})\|^2]\right\} \\ &\leq \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^3\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2], \end{aligned} \tag{29}$$

where the first inequality in Eq. (29) holds due to: 1) $\tau_t^i \leq \tau_{max}$ and 2) for a positive sequence $\boldsymbol{a}_t$, $\sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\boldsymbol{a}_u \leq \tau_{\max}(1-\beta_1)\sum_{t=1}^{T}\sum_{u=1}^{t}\beta_1^{t-u}\boldsymbol{a}_u \leq \tau_{\max}\sum_{t=1}^{T}\boldsymbol{a}_t$. In details,

$$\begin{aligned} &\sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\boldsymbol{a}_u \\ &= \sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)(\beta_1^{s-1}\boldsymbol{a}_1+\beta_1^{s-2}\boldsymbol{a}_2+\cdots+\beta_1^0\boldsymbol{a}_s) \\ &= \sum_{t=1}^{T}(1-\beta_1)\left[\sum_{s=t-\tau_t^i}^{t-1}\beta_1^{s-1}\boldsymbol{a}_1+\sum_{s=t-\tau_t^i}^{t-1}\beta_1^{s-2}\boldsymbol{a}_2+\cdots+\sum_{s=t-\tau_t^i}^{t-1}\beta_1^0\boldsymbol{a}_s\right] \\ &\leq \tau_{\max}\sum_{t=1}^{T}(1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\boldsymbol{a}_u \\ &\leq \tau_{\max}\sum_{t=1}^{T}\boldsymbol{a}_t. \end{aligned} \tag{30}$$

The second inequality in Eq. (29) hold by the fact of $\sum_{t=1}^{T}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^j})\|^2] \leq \tau_{\max}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]$. Similar, for term A_4, we have

$$A_4 = \frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}(1-\beta_1)\sum_{u=1}^{s}\beta_1^{s-u}\frac{M(M-1)}{N(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_t^j,k})\right\|^2\right]\Bigg\}$$

$$\leq \frac{\eta^3\eta_l^3 KL^2}{M^2\epsilon^3}\tau_{\max}^2 \frac{2M(M-1)}{N(N-1)}\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_t^j,k})\Big\|^2\Big]\Big\}, \tag{31}$$

With the term of A_0 to A_4, by Lemma C.3 and Lemma C.4, we have the following for Eq. (25),

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{T+1}) - f(\boldsymbol{z}_1)] \\
\leq& -\frac{\eta\eta_l K}{2C_G}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta\eta_l KL^2}{\epsilon}\Big[5K\eta_l^2 T(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\Big] \\
&+ \frac{2\eta^3\eta_l^3K^2L^2\tau_{\text{avg}}\tau_{\max}T}{M\epsilon^3}\sigma^2 + \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^3\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2 + K^2)\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&+ \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\cdot\frac{6M(N-M)}{N-1}[5K^3L^2\eta_l^2(\sigma^2 + 6K\sigma_g^2) + K^2\sigma_g^2]\cdot T\tau_{\text{avg}}\tau_{\max} \\
&+ \left(\frac{\beta_1}{1-\beta_1}\eta\eta_l KG^2 + \frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta^2\eta_l^2K^2G^2L\right)\frac{d}{\epsilon} + \frac{\beta_1^2}{(1-\beta_1)^2}\eta^2\eta_l^2K^2G^2L\frac{d}{\epsilon^2} \\
&+ \left(\frac{3\eta^2L}{2\epsilon^2} + \frac{\eta^2L}{2\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\sum_{t=1}^{T}\Big\{\frac{2\eta_l^2K}{M}\sigma^2 + \frac{2\eta_l^2(N-M)}{NM(N-1)}\Big[15NK^3L^2\eta_l^2(\sigma^2 + 6K\sigma_g^2) + (90K^4L^2\eta_l^2 + 3K^2) \\
&\cdot\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] + 3NK^2\sigma_g^2\Big] + \frac{2\eta_l^2(M-1)}{NM(N-1)}\mathbb{E}\Big[\Big\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big]\Big\} \\
&+ \frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^2\frac{M(M-1)}{N(N-1)}\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_t^j,k})\Big\|^2\Big]\Big\} \\
&- \frac{\eta\eta_l}{2KN^2C_G}\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big],
\end{aligned} \tag{32}$$

thus

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{T+1}) - f(\boldsymbol{z}_1)] \\
\leq& -\frac{\eta\eta_l K}{2C_G}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta\eta_l KL^2}{\epsilon}\Big[5K\eta_l^2T(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\tau_{\max}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]\Big] \\
&+ \frac{2\eta^3\eta_l^3K^2L^2\tau_{\text{avg}}\tau_{\max}T}{M\epsilon^3}\sigma^2 + \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^3\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2 + K^2)\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&+ \frac{\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\frac{6M(N-M)}{N-1}[5K^3L^2\eta_l^2(\sigma^2 + 6K\sigma_g^2) + K^2\sigma_g^2]\cdot T\tau_{\text{avg}}\tau_{\max} \\
&+ \left(\frac{\beta_1}{1-\beta_1}\eta\eta_l KG^2 + \frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta^2\eta_l^2K^2G^2L\right)\frac{d}{\epsilon} + \frac{\beta_1^2}{(1-\beta_1)^2}\eta^2\eta_l^2K^2G^2L\frac{d}{\epsilon^2} \\
&+ \left(\frac{3\eta^2L}{\epsilon^2} + \frac{\eta^2L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\sum_{t=1}^{T}\Big\{\frac{K\eta_l^2}{M}\sigma^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}\Big[15NK^3L^2\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 3NK^2\sigma_g^2\Big]\Big\} \\
&+ \left(\frac{3\eta^2L}{\epsilon^2} + \frac{\eta^2L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\frac{\eta_l^2(N-M)}{M(N-1)}(90K^4L^2\eta_l^2 + 3K^2)N\tau_{\max}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&+ \Big[\left(\frac{3\eta^2L}{\epsilon^2} + \frac{\eta^2L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\frac{\eta_l^2(M-1)}{NM(N-1)} + \frac{2\eta^3\eta_l^3KL^2}{M^2\epsilon^3}\tau_{\max}^2\frac{M(M-1)}{N(N-1)} - \frac{\eta\eta_l}{2KN^2C_G}\Big] \\
&\cdot\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big].
\end{aligned} \tag{33}$$

If the learning rates satisfy $\eta_l \leq \frac{1}{8KL}$ and

$$\begin{aligned}
&\eta\eta_l \leq \frac{\epsilon^2 M(N-1)}{4C_G N(M-1)KL}\left(3+\frac{\beta_1^2}{(1-\beta_1)^2}\right)^{-1}, \eta\eta_l \leq \frac{\sqrt{\epsilon^3 M(N-1)}}{\sqrt{8C_G N(M-1)}}\frac{1}{L\tau_{\max}},\\
&\eta_l \leq \frac{\sqrt{\epsilon}}{\sqrt{360C_G\tau_{\max}}KL}, \eta\eta_l \leq \frac{\epsilon^2 M(N-1)}{60C_G N(N-M)KL\tau_{\max}}\left(3+\frac{\beta_1^2}{(1-\beta_1)^2}\right)^{-1},\\
&\eta\eta_l \leq \frac{\sqrt{\epsilon^3 M(N-1)}}{12\sqrt{C_G N(M-1)\tau_{\max}^3}KL},
\end{aligned} \tag{34}$$

then we have

$$\begin{aligned}
&\left(\frac{3\eta^2 L}{\epsilon^2}+\frac{\eta^2 L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\frac{\eta_l^2(M-1)}{NM(N-1)}+\frac{2\eta^3\eta_l^3 KL^2}{M^2\epsilon^3}\tau_{\max}^2\frac{M(M-1)}{N(N-1)}-\frac{\eta\eta_l}{2KN^2C_G}\leq 0\\
&\frac{\eta\eta_l KL^2}{\epsilon}30K^2\eta_l^2\tau_{\max}+\left(\frac{3\eta^2 L}{\epsilon^2}+\frac{\eta^2 L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\frac{\eta_l^2(N-M)}{M(N-1)}(90K^4L^2\eta_l^2+3K^2)N\tau_{\max}\\
&\quad+\frac{\eta^3\eta_l^3 KL^2}{M^2\epsilon^3}\tau_{\max}^3\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\leq\frac{\eta\eta_l K}{4C_G}.
\end{aligned} \tag{35}$$

Thus Eq. (33) becomes

$$\begin{aligned}
\frac{\sum_{t=1}^T\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T}\leq&\frac{4C_G}{\eta\eta_l KT}[f(\boldsymbol{z}_1)-\mathbb{E}[f(\boldsymbol{z}_{T+1})]]+20C_G\epsilon^{-1}L^2K\eta_l^2(\sigma^2+6K\sigma_g^2)+\frac{8C_G\eta^2\eta_l^2KL^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\sigma^2\\
&+\frac{24C_G\eta^2\eta_l^2L^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\frac{N-M}{N-1}\cdot[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\\
&+\left(\frac{\beta_1}{1-\beta_1}G^2+\frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta\eta_l KG^2L\right)\frac{4C_Gd}{T\epsilon}+\frac{\beta_1^2}{(1-\beta_1)^2}\eta\eta_l KG^2L\frac{4C_Gd}{T\epsilon^2}\\
&+4C_G\left(\frac{3\eta L}{\epsilon^2}+\frac{\eta L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\left\{\frac{\eta_l}{M}\sigma^2+\frac{\eta_l(N-M)}{M(N-1)}[15K^2L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+3K\sigma_g^2]\right\}.
\end{aligned} \tag{36}$$

With $C_G=\eta_l KG+\epsilon$, Eq. (36) becomes

$$\begin{aligned}
&\frac{\sum_{t=1}^T\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T}\\
&\leq\frac{4(\eta_l KG+\epsilon)}{\eta\eta_l KT}[f(\boldsymbol{z}_1)-\mathbb{E}[f(\boldsymbol{z}_{t+1})]]+20(\eta_l KG+\epsilon)\epsilon^{-1}L^2K\eta_l^2(\sigma^2+6K\sigma_g^2)\\
&\quad+\frac{8(\eta_l KG+\epsilon)\eta^2\eta_l^2KL^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\sigma^2+\frac{24(\eta_l KG+\epsilon)\eta^2\eta_l^2L^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\frac{N-M}{N-1}\\
&\quad\cdot[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\\
&\quad+\left(\frac{\beta_1}{1-\beta_1}G^2+\frac{\beta_1^2}{(1-\beta_1)^2\epsilon}\eta\eta_l KG^2L\right)\frac{4(\eta_l KG+\epsilon)d}{T\epsilon}+\frac{\beta_1^2}{(1-\beta_1)^2}\eta\eta_l KG^2L\frac{4(\eta_l KG+\epsilon)d}{T\epsilon^2}\\
&\quad+4(\eta_l KG+\epsilon)\left(\frac{3\eta L}{\epsilon^2}+\frac{\eta L}{\epsilon^2}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\left\{\frac{\eta_l}{M}\sigma^2+\frac{\eta_l(N-M)}{M(N-1)}[15K^2L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+3K\sigma_g^2]\right\}.
\end{aligned} \tag{37}$$

For $\beta_1=0$, with the definition of $\mathcal{F}=f(\boldsymbol{x}_1)-\min_{\boldsymbol{x}}f(\boldsymbol{x})$, we have the following bound

$$\begin{aligned}
&\frac{\sum_{t=1}^T\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T}\\
&\leq\frac{4(\eta_l KG+\epsilon)}{\eta\eta_l KT}[f(\boldsymbol{z}_1)-\mathbb{E}[f(\boldsymbol{z}_{t+1})]]+20(\eta_l KG+\epsilon)\epsilon^{-1}L^2K\eta_l^2(\sigma^2+6K\sigma_g^2)\\
&\quad+\frac{8(\eta_l KG+\epsilon)\eta^2\eta_l^2KL^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\sigma^2+\frac{24(\eta_l KG+\epsilon)\eta^2\eta_l^2KL^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\frac{N-M}{N-1}
\end{aligned}$$

$$\cdot [5\eta_l^2 K^2 L^2(\sigma^2 + 6K\sigma_g^2) + K\sigma_g^2]$$
$$+ 12(\eta_l KG + \epsilon)\frac{\eta L}{\epsilon^2}\left\{\frac{\eta_l}{M}\sigma^2 + \frac{\eta_l(N-M)}{M(N-1)}[15\eta_l^2 K^2 L^2(\sigma^2 + 6K\sigma_g^2) + 3K\sigma_g^2]\right\}. \tag{38}$$

This concludes the proof. □

Proof of Corollary 5.7. From Eq. (38), we have the following bound

$$\begin{aligned}
&\frac{\sum_{t=1}^T \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T} \\
&= \mathcal{O}\Bigg(\frac{(\eta_l KG + \epsilon)}{\eta\eta_l KT}\mathcal{F} + (\eta_l KG + \epsilon)\frac{\eta_l^2 KL^2(\sigma^2 + K\sigma_g^2)}{\epsilon} \\
&\quad + \frac{(\eta_l KG + \epsilon)\eta^2\eta_l^2 KL^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\sigma^2 + \frac{(\eta_l KG + \epsilon)\eta^2\eta_l^2 L^2\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\frac{N-M}{N-1}[K^3L^2\eta_l^2(\sigma^2 + K\sigma_g^2) + K^2\sigma_g^2] \\
&\quad + (\eta_l KG + \epsilon)\frac{\eta\eta_l L}{M\epsilon^2}\left[\sigma^2 + \frac{N-M}{N-1}K\sigma_g^2\right] + (\eta_l KG + \epsilon)\frac{\eta\eta_l KL}{M\epsilon^2}\frac{N-M}{N-1}[\eta_l^2 KL^2(\sigma^2 + K\sigma_g^2)]\Bigg).
\end{aligned} \tag{39}$$

Reorganizing Eq. (39), particularly merging the stochastic variance and the global variance, we get

$$\begin{aligned}
&\frac{\sum_{t=1}^T \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T} \\
&= \mathcal{O}\Bigg(\frac{(\eta_l KG + \epsilon)}{\eta\eta_l KT}\mathcal{F} + (\eta_l KG + \epsilon)\frac{\eta_l^2 KL^2}{\epsilon}(\sigma^2 + K\sigma_g^2) \\
&\quad + (\eta_l KG + \epsilon)\frac{\eta^2\eta_l^2 KL^2}{M\epsilon^3}\left(\tau_{\text{avg}}\tau_{\max}\sigma^2 + \frac{N-M}{N-1}\tau_{\text{avg}}\tau_{\max}K\sigma_g^2\right) \\
&\quad + \frac{(\eta_l KG + \epsilon)\eta^2\eta_l^4 K^3L^4\tau_{\text{avg}}\tau_{\max}}{M\epsilon^3}\frac{N-M}{N-1}(\sigma^2 + K\sigma_g^2) \\
&\quad + (\eta_l KG + \epsilon)\frac{\eta\eta_l L}{M\epsilon^2}\left[\sigma^2 + \frac{N-M}{N-1}K\sigma_g^2\right] + (\eta_l KG + \epsilon)\frac{\eta\eta_l^3 K^2L^3}{M\epsilon^2}\frac{N-M}{N-1}(\sigma^2 + K\sigma_g^2)\Bigg).
\end{aligned} \tag{40}$$

By choosing $\eta = \Theta(\sqrt{M})$ and $\eta_l = \Theta(\frac{\sqrt{\mathcal{F}}}{\sqrt{TK(\sigma^2+K\sigma_g^2)L}})$, which implies $\eta\eta_l = \Theta(\frac{\sqrt{\mathcal{F}M}}{\sqrt{TK(\sigma^2+K\sigma_g^2)L}})$, and $\eta_l KG = \Theta(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}})$,

$$\begin{aligned}
&\frac{\sum_{t=1}^T \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T} \\
&= \mathcal{O}\Bigg\{\frac{\mathcal{F}G}{T\sqrt{M}} + \frac{\epsilon\sqrt{\mathcal{F}(\sigma^2 + K\sigma_g^2)L}}{\sqrt{TKM}} + \left(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2 + K\sigma_g^2)L}} + \epsilon\right)\frac{\mathcal{F}L}{T\epsilon} \\
&\quad + \left(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2 + K\sigma_g^2)L}} + \epsilon\right)\left(\frac{\mathcal{F}L\tau_{\text{avg}}\tau_{\max}}{T\epsilon^3} + \frac{N-M}{N-1}\frac{\mathcal{F}L\tau_{\text{avg}}\tau_{\max}}{T\epsilon^3}\right) \\
&\quad + \left(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2 + K\sigma_g^2)L}} + \epsilon\right)\left(\frac{\sqrt{\mathcal{F}L}\sigma}{\sqrt{TKM}} + \frac{N-M}{N-1}\frac{\sqrt{\mathcal{F}L}\sigma_g}{\sqrt{TM}}\right) + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2}\Bigg\}.
\end{aligned} \tag{41}$$

We again generalize terms with smaller T dependency orders, then we have

$$\begin{aligned}
&\frac{\sum_{t=1}^T \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{T} \\
&= \mathcal{O}\Bigg(\frac{\mathcal{F}G}{T\sqrt{M}} + \frac{\epsilon\sqrt{\mathcal{F}L}\sigma}{\sqrt{TKM}} + \frac{\epsilon\sqrt{\mathcal{F}L}\sigma_g}{\sqrt{TM}} + \frac{\mathcal{F}L}{T} + \frac{\mathcal{F}L\tau_{\text{avg}}\tau_{\max}}{T\epsilon^2} + \frac{N-M}{N-1}\frac{\mathcal{F}L\tau_{\text{avg}}\tau_{\max}}{T\epsilon^2}
\end{aligned}$$

$$+ \frac{\mathcal{F}G}{T\sqrt{M}} + \frac{\epsilon\sqrt{\mathcal{F}L}\sigma}{\sqrt{TKM}} + \frac{N-M}{N-1}\frac{\mathcal{F}G}{T\sqrt{M}} + \frac{N-M}{N-1}\frac{\epsilon\sqrt{\mathcal{F}L}\sigma_g}{\sqrt{TM}} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2}\Bigg)$$
$$= \mathcal{O}\Bigg(\frac{\sqrt{\mathcal{F}}\sigma}{\sqrt{TKM}} + \frac{\sqrt{\mathcal{F}}\sigma_g}{\sqrt{TM}} + \frac{\mathcal{F}}{T} + \frac{\mathcal{F}G}{T\sqrt{M}} + \frac{\mathcal{F}\tau_{\max}\tau_{\text{avg}}}{T}\Bigg). \tag{42}$$

This concludes the proof for Corollary 5.7. □

B. Convergence analysis for delay adaptive asynchronous FL

Proof of Theorem 5.9. For the proof of delay adaptive, for proof convenience, we conduct analysis under the case that $\beta_1 = 0$. From Assumption 5.1, f is L-smooth, then taking conditional expectation at time t on the auxiliary sequence $\boldsymbol{x}_t$, we have

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)] \\
&= \mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)] \\
&\leq \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] \\
&= \underbrace{\mathbb{E}\Bigg[\Bigg\langle \nabla f(\boldsymbol{x}_t), \eta_t \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}\Bigg\rangle\Bigg]}_{I_1} + \underbrace{\frac{\eta_t^2 L}{2}\mathbb{E}\Bigg[\Bigg\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}\Bigg\|^2\Bigg]}_{I_2}.
\end{aligned} \tag{43}$$

Bounding I_1 We have

$$\begin{aligned}
I_1 &= \eta_t \mathbb{E}\Bigg[\Bigg\langle \nabla f(\boldsymbol{x}_t), \frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}\Bigg\rangle\Bigg] \\
&= \eta_t \mathbb{E}\Bigg[\Bigg\langle \nabla f(\boldsymbol{x}_t), \frac{\bar{\boldsymbol{\Delta}}_t}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}\Bigg\rangle\Bigg] \\
&= \eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, \bar{\boldsymbol{\Delta}}_t + \eta_l K \nabla f(\boldsymbol{x}_t) - \eta_l K \nabla f(\boldsymbol{x}_t)\Bigg\rangle\Bigg] \\
&= -\eta_t \eta_l K \mathbb{E}\Bigg[\Bigg\|\frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}}\Bigg\|^2\Bigg] + \eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, \bar{\boldsymbol{\Delta}}_t + \eta_l K \nabla f(\boldsymbol{x}_t)\Bigg\rangle\Bigg] \\
&= -\eta_t \eta_l K \mathbb{E}\Bigg[\Bigg\|\frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}}\Bigg\|^2\Bigg] + \eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, -\frac{1}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1} \eta_l \boldsymbol{g}^i_{t-\tau^i_t,k} + \frac{\eta_l K}{N}\sum_{i\in[N]} \nabla F_i(\boldsymbol{x}_t)\Bigg\rangle\Bigg],
\end{aligned} \tag{44}$$

where $\bar{\boldsymbol{\Delta}}_t = -\frac{1}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1} \eta_l \boldsymbol{g}^i_{t-\tau^i_t,k}$. For the inner product term in (44), by the fact of $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}[\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a} - \boldsymbol{b}\|^2]$, we have

$$\begin{aligned}
&\eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\nabla f(\boldsymbol{x}_t)}{\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon}, -\frac{1}{N}\sum_{i\in[N]}\sum_{k=0}^{K-1} \eta_l \boldsymbol{g}^i_{t-\tau^i_t,k} + \frac{\eta_l K}{N}\sum_{i\in[N]} \nabla F_i(\boldsymbol{x}_t)\Bigg\rangle\Bigg] \\
&= \eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}} \nabla f(\boldsymbol{x}_t), -\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}} \frac{1}{NK}\sum_{i\in[N]}\sum_{k=0}^{K-1} (\boldsymbol{g}^i_{t-\tau^i_t,k} - \nabla F_i(\boldsymbol{x}_t))\Bigg\rangle\Bigg] \\
&= \eta_t \mathbb{E}\Bigg[\Bigg\langle \frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}} \nabla f(\boldsymbol{x}_t), -\frac{\sqrt{\eta_l K}}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}} \frac{1}{NK}\sum_{i\in[N]}\sum_{k=0}^{K-1} (\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\Bigg\rangle\Bigg] \\
&= \frac{\eta_t \eta_l K}{2}\mathbb{E}\Bigg[\Bigg\|\frac{\nabla f(\boldsymbol{x}_t)}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}}\Bigg\|^2\Bigg] + \frac{\eta_t \eta_l}{2N^2 K}\mathbb{E}\Bigg[\Bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t} + \epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1} (\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) - \nabla F_i(\boldsymbol{x}_t))\Bigg\|^2\Bigg]
\end{aligned}$$

$$-\frac{\eta_t\eta_l}{2N^2K}\mathbb{E}\bigg[\bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg], \tag{45}$$

where second equation holds by $\mathbb{E}[\boldsymbol{g}^i_{t-\tau^i_t,k}]=\mathbb{E}[\nabla F(\boldsymbol{x}^i_{t-\tau^i_t,k})]$, for the second term in Eq. (45), we have

$$\begin{aligned}
&\frac{\eta_t\eta_l}{2N^2K}\mathbb{E}\bigg[\bigg\|\frac{1}{(\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon)^{1/2}}\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})-\nabla F_i(\boldsymbol{x}_t))\bigg\|^2\bigg]\\
&\le\frac{\eta_t\eta_l}{2N^2K\epsilon}\mathbb{E}\bigg[\bigg\|\sum_{i\in[N]}\sum_{k=0}^{K-1}(\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})-\nabla F_i(\boldsymbol{x}_t))\bigg\|^2\bigg]\\
&\le\frac{\eta_t\eta_l}{2N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2]\\
&\le\frac{\eta_t\eta_l}{N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\bigg[\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\tau^i_t})\|^2]+\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_{t-\tau^i_t})-\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2]\bigg]\\
&\le\frac{\eta_t\eta_l}{N\epsilon}\sum_{i\in[N]}\sum_{k=0}^{K-1}\bigg[L^2\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}_{t-\tau^i_t}\|^2]+L^2\mathbb{E}[\|\boldsymbol{x}_{t-\tau^i_t}-\boldsymbol{x}^i_{t-\tau^i_t,k}\|^2]\bigg].
\end{aligned}\tag{46}$$

where the first second inequality holds by $\forall\boldsymbol{a}_i,\|\sum_{i=1}^n\boldsymbol{a}_i\|^2\le n\sum_{i=1}^n\|\boldsymbol{a}_i\|^2$, and the last inequality holds by Assumption 5.1. For the second term in Eq. (46), following by Lemma C.5, there is

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{t-\tau^i_t}-\boldsymbol{x}^i_{t-\tau^i_t,k}\|^2]&=\mathbb{E}\bigg[\bigg\|\sum_{m=0}^{k-1}\eta_l\boldsymbol{g}^i_{t-\tau^i_t,m}\bigg\|^2\bigg]\\
&\le 5K\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2].
\end{aligned}\tag{47}$$

For the first term in Eq. (46), we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}_{t-\tau^i_t}\|^2]&=\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}(\boldsymbol{x}_{s+1}-\boldsymbol{x}_s)\bigg\|^2\bigg]\\
&=\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\frac{\boldsymbol{\Delta}_s}{\sqrt{\widehat{\boldsymbol{v}}_s}+\epsilon}\bigg\|^2\bigg]\\
&\le\frac{1}{\epsilon^2}\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\boldsymbol{\Delta}_s\bigg\|^2\bigg]\\
&=\frac{1}{\epsilon^2}\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\frac{1}{M}\sum_{j\in\mathcal{M}_s}\boldsymbol{\Delta}^j_s\bigg\|^2\bigg]\\
&=\frac{1}{\epsilon^2}\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\frac{1}{M}\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\eta_l\boldsymbol{g}^j_{s-\tau^j_s,k}\bigg\|^2\bigg],
\end{aligned}\tag{48}$$

then by decomposing stochastic noise, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}_{t-\tau^i_t}\|^2]\\
&\le\frac{1}{\epsilon^2}\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\frac{1}{M}\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\eta_l[\boldsymbol{g}^j_{s-\tau^j_s,k}-\nabla F_j(\boldsymbol{x}^j_{s-\tau^j_s,k})+\nabla F_j(\boldsymbol{x}^j_{s-\tau^j_s,k})]\bigg\|^2\bigg]\\
&\le\frac{2}{\epsilon^2}\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}\eta_s\frac{1}{M}\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\eta_l[\boldsymbol{g}^j_{s-\tau^j_s,k}-\nabla F_j(\boldsymbol{x}^j_{s-\tau^j_s,k})]\bigg\|^2\bigg]
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2}{\epsilon^2}\mathbb{E}\Big[\Big\|\sum_{s=t-\tau_t^i}^{t-1}\eta_s\frac{1}{M}\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\eta_l\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big] \\
&\le \frac{2}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\frac{K\eta_l^2}{M}\sigma^2 + \frac{2\tau_t^i}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\frac{\eta_l^2}{M^2}\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big] \\
&\le \frac{2}{\epsilon^2}\tau_t^i\eta^2\frac{K\eta_l^2}{M}\sigma^2 + \frac{2\tau_t^i}{\epsilon^2}\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\frac{\eta_l^2}{M^2}\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big],
\end{aligned}
\tag{49}
$$

where the second inequality holds by $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \le 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$. The second inequality holds by Assumption 5.2, i.e., the zero-mean and the independency of stochastic noise. The last inequality in Eq. (49) holds due to the following: with adaptive learning rates

$$
\eta_t = \begin{cases} \eta & \text{if } \tau_t^{\max} \le \tau_c, \\ \min\{\eta, \frac{1}{\tau_t^{\max}}\} & \text{if } \tau_t^{\max} > \tau_c, \end{cases}
\tag{50}
$$

thus we have $\eta_s \le \eta$ in the last inequality in Eq. (49). Then for I_1, following Lemma C.2 $\frac{1}{C_G}\|\boldsymbol{x}\| \le \big\|\frac{\boldsymbol{x}}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\big\| \le \frac{1}{\epsilon}\|\boldsymbol{x}\|$ and $C_G = \eta_l KG + \epsilon$, we have

$$
\begin{aligned}
I_1 \le& -\frac{\eta_t\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta_t\eta_l}{2KC_G}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big] \\
&+ \frac{\eta_t\eta_l KL^2}{\epsilon}\Big[5K\eta_l^2(\sigma^2+6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\Big] + \frac{2\eta_t\eta^2\eta_l^3K^2L^2}{M\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau_t^i\sigma^2 \\
&+ \frac{2\eta_t\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big].
\end{aligned}
\tag{51}
$$

Bounding I_2

$$
I_2 = \frac{\eta_t^2 L}{2}\mathbb{E}\Big[\Big\|\frac{\boldsymbol{\Delta}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\Big\|^2\Big] \le \frac{\eta_t^2 L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2],
\tag{52}
$$

where the first inequality follows by Cauchy-Schwarz inequality.

Merging pieces. Therefore, by merging pieces together, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)] = I_1 + I_2 \\
\le& -\frac{\eta_t\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta_t\eta_l}{2KC_G}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})\Big\|^2\Big] \\
&+ \frac{\eta_t\eta_l KL^2}{\epsilon}\Big[5K\eta_l^2(\sigma^2+6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\Big] + \frac{2\eta_t\eta^2\eta_l^3K^2L^2}{M\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau_t^i\sigma^2 \\
&+ \frac{2\eta_t\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big] + \frac{\eta_t^2 L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2].
\end{aligned}
\tag{53}
$$

Denote a sequences $\boldsymbol{G}_s = \sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_s^j,k})$, then re-write and organize the above inequality, we have

$$
\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)]
$$

$$
\begin{aligned}
&\leq -\frac{\eta_t\eta_l K}{2C_G}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta_t\eta_l}{2KC_G}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\right\|^2\right] \\
&\quad + \frac{\eta_t\eta_l K L^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2) + 30K^2\eta_l^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]\right] + \frac{2\eta_t\eta^2\eta_l^3K^2L^2}{M\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau^i_t\sigma^2 \\
&\quad + \frac{2\eta_t\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\mathbb{E}[\|\boldsymbol{G}_s\|^2] + \frac{\eta_t^2L}{2\epsilon^2}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2]. \qquad (54)
\end{aligned}
$$

Summing over $t=1$ to T, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{T+1}) - f(\boldsymbol{x}_1)] \\
&\leq -\frac{\eta_l K}{2C_G}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta_l KL^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2)\sum_{t=1}^{T}\eta_t + 30K^2\eta_l^2\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]\right] \\
&\quad + \frac{2\eta^2\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\tau^i_t\eta_t + \frac{2\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\mathbb{E}[\|\boldsymbol{G}_s\|^2] + \frac{L}{2\epsilon^2}\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\quad - \frac{\eta_l}{2KC_G}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\right\|^2\right] \\
&\leq -\frac{\eta_l K}{2C_G}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta_l KL^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2)\sum_{t=1}^{T}\eta_t + 30K^2\eta_l^2\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]\right] \\
&\quad + \frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}\sigma^2T\tau_{\text{avg}} + \underbrace{\frac{2\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\mathbb{E}[\|\boldsymbol{G}_s\|^2]}_{A_1} + \frac{L}{2\epsilon^2}\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\quad - \frac{\eta_l}{2KC_G}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\right\|^2\right], \qquad (55)
\end{aligned}
$$

where the second inequality holds by $\eta_t \leq \eta$. We have the following for term A_1,

$$
\begin{aligned}
&\frac{2\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\mathbb{E}[\|\boldsymbol{G}_s\|^2] \\
&= \frac{2\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\cdot\frac{3M(N-M)}{N-1} \\
&\quad\cdot\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2) + (30K^4L^2\eta_l^2+K^2)\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau^j_s})\|^2] + K^2\sigma_g^2\right] \\
&\quad + \frac{2\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\cdot\frac{M(M-1)}{n(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau^j_s,k})\right\|^2\right] \\
&= \underbrace{\frac{\eta_l^3KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\cdot\frac{6M(N-M)}{N-1}\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2\right]}_{A_2} \\
&\quad + \underbrace{\frac{\eta_l^3KL}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau^i_t\sum_{s=t-\tau^i_t}^{t-1}\eta_s^2\cdot\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau^j_s})\|^2]}_{A_3}
\end{aligned}
$$

$$+ \underbrace{\frac{2\eta_l^3 KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\cdot\frac{M(M-1)}{N(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\right\|^2\right]}_{A_4}$$

$$= A_2 + A_3 + A_4. \tag{56}$$

For term A_2, note that with $\eta \le \frac{\sqrt{M}}{\tau_c}$, we have the adaptive learning rates

$$\eta_t = \begin{cases} \eta & \text{if } \tau_t^{\max} \le \tau_c, \\ \min\{\eta, \frac{1}{\tau_t^{\max}}\} & \text{if } \tau_t^{\max} > \tau_c, \end{cases} \tag{57}$$

which implies that $\eta_t \le \eta$ and $\eta_t \le \min\{\frac{1}{\tau_t^{\max}}, \frac{\sqrt{M}}{\tau_c}\}$. Moreover, recall that $\tau_t^{\max} = \max_{i\in[N]}\{\tau_t^i\}$, for each i, we have $\eta_t\tau_t^i \le \frac{\sqrt{M}\cdot\tau_t^i}{\max(\tau_t^{\max},\tau_c)} \le \sqrt{M}$. by the fact of $\eta_t\tau_t^i \le \sqrt{M}$, $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_t^i \le T\tau_{\text{avg}}$ and $\eta_s \le \eta$, we have

$$\begin{aligned} A_2 &\le \frac{\eta_l^3 KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau_t^i\cdot\tau_t^i\eta^2\cdot\frac{6M(N-M)}{N-1}\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2\right] \\ &\le \frac{\eta_l^3 KL^2}{M\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sqrt{M}\cdot\tau_t^i\eta^2\cdot\frac{N-M}{N-1}\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2\right] \\ &\le \frac{\eta^2\eta_l^3 KL^2}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}\left[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2\right]T\tau_{\text{avg}}. \end{aligned} \tag{58}$$

For term A_3, since $\eta_s \le \eta$, and consider $\tau_t^i \le \tau_{\max}$ and $\sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}\boldsymbol{a}_s \le \tau_{\max}\sum_{t=1}^{T}\boldsymbol{a}_t$, we have

$$\begin{aligned} A_3 &\le \frac{\eta_l^3 KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\cdot\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2] \\ &\le \frac{\eta^2\eta_l^3 KL}{M^2\epsilon^3}\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\tau_{\max}^3\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]. \end{aligned} \tag{59}$$

For term A_4, similar to the proof of non-delay adaptive FADAS, by $\eta_s \le \eta$, $\tau_t^i \le \tau_{\max}$ and $\sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}\boldsymbol{a}_s \le \tau_{\max}\sum_{t=1}^{T}\boldsymbol{a}_t$, there is

$$\begin{aligned} A_4 &= \frac{2\eta_l^3 KL^2}{M^2\epsilon^3}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\eta_t\tau_t^i\sum_{s=t-\tau_t^i}^{t-1}\eta_s^2\cdot\frac{M(M-1)}{N(N-1)}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\right\|^2\right] \\ &\le \frac{2\eta^2\eta_l^3 KL^2}{M^2\epsilon^3}\frac{M(M-1)}{N(N-1)}\tau_{\max}\sum_{t=1}^{T}\eta_t\sum_{s=t-\tau_t^i}^{t-1}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\right\|^2\right]\Bigg\} \\ &\le \frac{2\eta^2\eta_l^3 KL^2}{M^2\epsilon^3}\frac{M(M-1)}{N(N-1)}\tau_{\max}^2\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{t-\tau_t^j,k})\right\|^2\right]\Bigg\}. \end{aligned} \tag{60}$$

By Lemma C.3 and Lemma C.4, we have the following for Eq. (55),

$$\begin{aligned} &\mathbb{E}[f(\boldsymbol{x}_{T+1}) - f(\boldsymbol{x}_1)] \\ \le& -\frac{\eta_l K}{2C_G}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta_l KL^2}{\epsilon}\left[5K\eta_l^2(\sigma^2+6K\sigma_g^2)\sum_{t=1}^{T}\eta_t + 30K^2\eta_l^2\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\right] \\ &+ \frac{2\eta^3\eta_l^3K^2L^2}{M\epsilon^3}T\tau_{\text{avg}}\sigma^2 + \frac{\eta^2\eta_l^3KL^2}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]T\tau_{\text{avg}} \end{aligned}$$

$$
\begin{aligned}
&+ \frac{\eta^2\eta_l^3 KL^2}{M^2\epsilon^3}\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\tau_{\max}^3\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&+ \frac{L}{2\epsilon^2}\sum_{t=1}^{T}\eta_t^2\Bigg\{\frac{2K\eta_l^2}{M}\sigma^2+\frac{2\eta_l^2(N-M)}{NM(N-1)}\Big[15NK^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+3NK^2\sigma_g^2\Big] \\
&+ \frac{2\eta_l^2(N-M)}{M(N-1)}(90NK^4L^2\eta_l^2+3NK^2)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\Bigg\} \\
&+ \left[\frac{\eta L}{\epsilon^2}\frac{\eta_l^2(M-1)}{NM(N-1)}+\frac{2\eta^2\eta_l^3KL^2}{M^2\epsilon^3}\frac{M(M-1)}{N(N-1)}\tau_{\max}^2-\frac{\eta_l}{2KN^2C_G}\right]\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i)\right\|^2\right],
\end{aligned}
\tag{61}
$$

by the relationship of $\sum_{t=1}^{T}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]\le\tau_{\max}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]$.

If the learning rates satisfy $\eta_l\le\frac{1}{8KL}$ and

$$
\begin{aligned}
&\eta\eta_l\le\frac{\epsilon^2M(N-1)}{4C_GN(M-1)KL},\eta\eta_l\le\frac{\sqrt{\epsilon^3M(N-1)}}{\sqrt{8C_GN(M-1)}}\frac{1}{L\tau_{\max}}, \\
&\eta_l\le\frac{\sqrt{\epsilon}}{\sqrt{360C_G\tau_{\max}}KL},\eta\eta_l\le\frac{\epsilon^2M(N-1)}{60C_GN(N-M)KL\tau_{\max}}, \\
&\eta\eta_l\le\frac{\sqrt{\epsilon^3M(N-1)}}{12\sqrt{C_GN(M-1)\tau_{\max}^3}KL}.
\end{aligned}
\tag{62}
$$

Then we have

$$
\begin{aligned}
&\frac{\eta L}{\epsilon^2}\frac{\eta_l^2(M-1)}{NM(N-1)}+\frac{2\eta^2\eta_l^3KL^2}{M^2\epsilon^3}\frac{M(M-1)}{N(N-1)}\tau_{\max}^2-\frac{\eta_l}{2KN^2C_G}\le 0 \\
&\frac{\eta_lKL^2}{\epsilon}30K^2\eta_l^2\tau_{\max}+\frac{\eta L}{\epsilon^2}\frac{\eta_l^2(N-M)}{M(N-1)}(90K^4L^2\eta_l^2+3K^2)N\tau_{\max} \\
&\quad+\frac{\eta^2\eta_l^3KL^2}{M^2\epsilon^3}\frac{6M(N-M)}{N-1}(30K^4L^2\eta_l^2+K^2)\tau_{\max}^3\le\frac{\eta_lK}{4C_G}.
\end{aligned}
\tag{63}
$$

Thus Eq. (61) becomes

$$
\begin{aligned}
&\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
\le&\frac{4C_G}{\eta_lK}[f(\boldsymbol{x}_1)-\mathbb{E}[f(\boldsymbol{x}_{T+1})]]+\frac{4C_GL^2}{\epsilon}5K\eta_l^2(\sigma^2+6K\sigma_g^2)\sum_{t=1}^{T}\eta_t \\
&+\frac{8C_G\eta^3\eta_l^2KL^2}{M\epsilon^3}T\tau_{\text{avg}}\sigma^2+\frac{24C_G\eta^2\eta_l^2L^2}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]T\tau_{\text{avg}} \\
&+\frac{4C_G\eta L}{\epsilon^2}\sum_{t=1}^{T}\eta_t\left[\frac{\eta_l}{M}\sigma^2+\frac{\eta_l(N-M)}{NM(N-1)}[15NK^2L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+3NK\sigma_g^2]\right],
\end{aligned}
\tag{64}
$$

divided by the learning rates,

$$
\begin{aligned}
\frac{\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{\sum_{t=1}^{T}\eta_t}\le&\frac{4C_G}{\eta_lK\sum_{t=1}^{T}\eta_t}[f(\boldsymbol{x}_1)-\mathbb{E}[f(\boldsymbol{x}_{T+1})]] \\
&+20C_G\epsilon^{-1}L^2K\eta_l^2(\sigma^2+6K\sigma_g^2)+\frac{8C_G\eta^3\eta_l^2KL^2}{M\epsilon^3}\frac{T\tau_{\text{avg}}}{\sum_{t=1}^{T}\eta_t}\sigma^2 \\
&+\frac{24C_G\eta^2\eta_l^2L^2}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}[5K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+K^2\sigma_g^2]\frac{T\tau_{\text{avg}}}{\sum_{t=1}^{T}\eta_t}
\end{aligned}
$$

$$+ \frac{4C_G \eta L}{\epsilon^2}\bigg\{\frac{\eta_l}{M}\sigma^2 + \frac{\eta_l(N-M)}{M(N-1)}[15K^2TL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+3K\sigma_g^2]\bigg\}. \tag{65}$$

This concludes the proof. □

Proof of Corollary 5.10. Since the delay adaptive learning rate satisfy, $\eta_t \le \eta$, and when $\tau_c = \tau_{\text{median}}$, there is $\sum_{t=1}^T \eta_t \ge \sum_{t:\tau_t \le \tau_c} \eta \ge \frac{T\eta}{2}$ (since there are at least half of the iterations with the delay smaller than τ_c). Recalling that $C_G = \eta_l KG + \epsilon$, then

$$\begin{aligned}
&\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{\sum_{t=1}^T \eta_t} \\
&= \mathcal{O}\bigg\{\frac{(\eta_l KG+\epsilon)}{\eta\eta_l KT}\mathcal{F} + (\eta_l KG+\epsilon)\frac{\eta_l^2 KL^2(\sigma^2+K\sigma_g^2)}{\epsilon} \\
&\quad + \frac{(\eta_l KG+\epsilon)\eta^2\eta_l^2 KL^2\tau_{\text{avg}}}{M\epsilon^3}\sigma^2 + \frac{(\eta_l KG+\epsilon)\eta\eta_l^2 KL^2\tau_{\text{avg}}}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}[K^2L^2\eta_l^2(\sigma^2+K\sigma_g^2)+K\sigma_g^2] \\
&\quad + (\eta_l KG+\epsilon)\frac{\eta\eta_l L}{M\epsilon^2}\left[\sigma^2 + \frac{N-M}{N-1}K\sigma_g^2\right] + (\eta_l KG+\epsilon)\frac{\eta\eta_l KL}{M\epsilon^2}\frac{N-M}{N-1}[\eta_l^2 KL^2(\sigma^2+K\sigma_g^2)]\bigg\}.
\end{aligned} \tag{66}$$

Reorganizing Eq. (66), particularly merging the stochastic variance and the global variance, then we have

$$\begin{aligned}
&\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{\sum_{t=1}^T \eta_t} \\
&= \mathcal{O}\bigg\{\frac{(\eta_l KG+\epsilon)}{\eta\eta_l KT}\mathcal{F} + (\eta_l KG+\epsilon)\frac{\eta_l^2 KL^2}{\epsilon}(\sigma^2+K\sigma_g^2) \\
&\quad + (\eta_l KG+\epsilon)\frac{\eta\eta_l^2 KL^2}{M\epsilon^3}\left(\eta\tau_{\text{avg}}\sigma^2 + \frac{\sqrt{M}(N-M)}{N-1}\tau_{\text{avg}}K\sigma_g^2\right) + \frac{(\eta_l KG+\epsilon)\eta\eta_l^4 K^3L^4\tau_{\text{avg}}}{\sqrt{M}\epsilon^3}\frac{N-M}{N-1}(\sigma^2+K\sigma_g^2) \\
&\quad + (\eta_l KG+\epsilon)\frac{\eta\eta_l L}{M\epsilon^2}\left[\sigma^2 + \frac{N-M}{N-1}K\sigma_g^2\right] + (\eta_l KG+\epsilon)\frac{\eta\eta_l^3 K^2L^3}{M\epsilon^2}\frac{N-M}{N-1}(\sigma^2+K\sigma_g^2)\bigg\}.
\end{aligned} \tag{67}$$

By choosing $\eta = \sqrt{M}/\tau_c$ and $\eta_l = \min\left\{\frac{1}{KL}, \frac{\tau_c\sqrt{\mathcal{F}}}{\sqrt{TK(\sigma^2+K\sigma_g^2)L}}\right\}$, which implies $\eta\eta_l = \min\left\{\frac{\sqrt{M}}{\tau_c KL}, \frac{\sqrt{\mathcal{F}M}}{\sqrt{TK(\sigma^2+K\sigma_g^2)L}}\right\}$,

$$\begin{aligned}
&\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{\sum_{t=1}^T \eta_t} \\
&= \mathcal{O}\bigg\{\left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\frac{\sqrt{\mathcal{F}(\sigma^2+K\sigma_g^2)L}}{\sqrt{TKM}} + \left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\frac{\mathcal{F}L\tau_c^2}{T\epsilon} \\
&\quad + \left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\left(\frac{\mathcal{F}L\tau_{\text{avg}}}{T\epsilon^3} + \frac{N-M}{N-1}\frac{\mathcal{F}L\tau_c\tau_{\text{avg}}}{T\epsilon^3}\right) \\
&\quad + \left(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\left(\frac{\sqrt{\mathcal{F}L}\sigma}{\sqrt{TKM}} + \frac{N-M}{N-1}\frac{\sqrt{\mathcal{F}L}\sigma_g}{\sqrt{TM}}\right) + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2}\bigg\} \\
&= \mathcal{O}\bigg\{\left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\frac{\sqrt{\mathcal{F}(\sigma^2+K\sigma_g^2)L}}{\sqrt{TKM}} + \left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\frac{\mathcal{F}L\tau_c^2}{T\epsilon} \\
&\quad + \left(\frac{\tau_c\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\left(\frac{\mathcal{F}L\tau_{\text{avg}}}{T\epsilon^3} + \frac{\mathcal{F}L\tau_c\tau_{\text{avg}}}{T\epsilon^3}\right) + \left(\frac{\sqrt{\mathcal{F}K}G}{\sqrt{T(\sigma^2+K\sigma_g^2)L}}+\epsilon\right)\left(\frac{\sqrt{\mathcal{F}L}\sigma}{\sqrt{TKM}} + \frac{\sqrt{\mathcal{F}L}\sigma_g}{\sqrt{TM}}\right) \\
&\quad + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2}\bigg\}.
\end{aligned} \tag{68}$$

We again generalize terms with smaller T dependency orders, then we have

$$\frac{\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]}{\sum_{t=1}^{T}\eta_t} \leq \mathcal{O}\left(\frac{\tau_c\mathcal{F}G}{T\sqrt{M}}+\frac{\sqrt{\mathcal{F}}\sigma}{\sqrt{TKM}}+\frac{\sqrt{\mathcal{F}}\sigma_g}{\sqrt{TM}}+\frac{\mathcal{F}\tau_c^2}{T}+\frac{\mathcal{F}\tau_{\text{avg}}}{T}+\frac{\mathcal{F}\tau_c\tau_{\text{avg}}}{T}\right), \tag{69}$$

reorganizing and then obtain the rate of convergence in Eq. (9).

□

C. Supporting Lemmas

Lemma C.1 (Lemma for momentum term in the update rule)**.** *The first order momentum terms $\boldsymbol{m}_t$ in Algorithm 1 hold the following relationship w.r.t. model difference $\boldsymbol{\Delta}_t$:*

$$\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{m}_t\|^2] \leq \sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2]. \tag{70}$$

Proof. By the updating rule, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{m}_t\|^2] &= \mathbb{E}\left[\left\|(1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\boldsymbol{\Delta}_u\right\|^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\boldsymbol{\Delta}_u^i\right)^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\right)\left(\sum_{u=1}^{t}\beta_1^{t-u}(\boldsymbol{\Delta}_u^i)^2\right)\right] \\
&\leq (1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\boldsymbol{\Delta}_u\|^2].
\end{aligned} \tag{71}$$

Summing over $t = 1, ..., T$ yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{m}_t\|^2] &= (1-\beta_1)\sum_{t=1}^{T}\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\boldsymbol{\Delta}_u\|^2] \\
&= (1-\beta_1)\sum_{u=1}^{T}\sum_{t=u}^{T}\beta_1^{t-u}\mathbb{E}[\|\boldsymbol{\Delta}_u\|^2] \\
&\leq (1-\beta_1)\sum_{u=1}^{T}\frac{1}{1-\beta_1}\mathbb{E}[\|\boldsymbol{\Delta}_u\|^2] \\
&= \sum_{u=1}^{T}\mathbb{E}[\|\boldsymbol{\Delta}_u\|^2].
\end{aligned} \tag{72}$$

This concludes the proof. □

Lemma C.2. *Under Assumptions 5.3, we have $\|\nabla f(\boldsymbol{x})\| \leq G$, $\|\boldsymbol{\Delta}_t\| \leq \eta_l KG$, $\|\boldsymbol{m}_t\| \leq \eta_l KG$, $\|\boldsymbol{v}_t\| \leq \eta_l^2K^2G^2$ and $\|\widehat{\boldsymbol{v}}_t\| \leq \eta_l^2K^2G^2$.*

Proof. Since f has G-bounded stochastic gradients, for any $\boldsymbol{x}$ and ξ, there is $\|\nabla f(\boldsymbol{x},\xi)\| \leq G$, thus it implies

$$\|\nabla f(\boldsymbol{x})\| = \|\mathbb{E}_\xi\nabla f(\boldsymbol{x},\xi)\| \leq \mathbb{E}_\xi\|\nabla f(\boldsymbol{x},\xi)\| \leq G.$$

For each model difference $\boldsymbol{\Delta}_t^i$ on client i, $\boldsymbol{\Delta}_t^i$ satisfies,

$$\boldsymbol{\Delta}_t^i = \boldsymbol{x}_{t,K}^i - \boldsymbol{x}_t = -\eta_l \sum_{k=0}^{K-1} \boldsymbol{g}_{t,k}^i,$$

therefore,

$$\|\boldsymbol{\Delta}_t^i\| = \left\| -\eta_l \sum_{k=0}^{K-1} \boldsymbol{g}_{t,k}^i \right\| \leq \eta_l K G,$$

for the global model difference $\boldsymbol{\Delta}_t$,

$$\|\boldsymbol{\Delta}_t\| = \left\| \frac{1}{M} \sum_{i \in \mathcal{M}_t} \boldsymbol{\Delta}_t^i \right\| \leq \eta_l K G.$$

Thus we can obtain the bound for momentum $\boldsymbol{m}_t$ and variance $\boldsymbol{v}_t$,

$$\|\boldsymbol{m}_t\| = \left\| (1-\beta_1) \sum_{s=1}^{t} \beta_1^{t-s} \boldsymbol{\Delta}_s \right\| \leq \eta_l K G, \quad \|\boldsymbol{v}_t\| = \left\| (1-\beta_2) \sum_{s=1}^{t} \beta_2^{t-s} \boldsymbol{\Delta}_s^2 \right\| \leq \eta_l^2 K^2 G^2.$$

By the updating rule of $\widehat{\boldsymbol{v}}_t$, there exists a $j \in [t]$ such that $\widehat{\boldsymbol{v}}_t = \boldsymbol{v}_j$. Then

$$\|\widehat{\boldsymbol{v}}_t\| \leq \eta_l^2 K^2 G^2. \tag{73}$$

This concludes the proof. □

Lemma C.3. *For the variance difference sequence* $\boldsymbol{V}_t = \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_t + \epsilon}} - \frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{t-1} + \epsilon}}$, *we have*

$$\sum_{t=1}^{T} \|\boldsymbol{V}_t\|_1 \leq \frac{d}{\epsilon}, \quad \sum_{t=1}^{T} \|\boldsymbol{V}_t\|_2^2 \leq \frac{d}{\epsilon^2}. \tag{74}$$

Proof. The proof of Lemma C.3 is exactly the same as the proof of Lemma C.2 in Wang et al. (2022b). □

Lemma C.4. *Recall the sequence* $\boldsymbol{\Delta}_t = \frac{1}{M} \sum_{i \in \mathcal{M}_t} \boldsymbol{\Delta}_{t-\tau_t^i}^i = -\frac{\eta_l}{M} \sum_{i \in \mathcal{M}_t} \sum_{k=0}^{K-1} \boldsymbol{g}_{t-\tau_t^i,k}^i = -\frac{\eta_l}{M} \sum_{i \in \mathcal{M}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i; \xi)$ *and* $\mathcal{M}_t$ *be the set that include client send the local updates to the server at global round* t. *The global model difference* $\boldsymbol{\Delta}_t$ *satisfies*

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] &= \mathbb{E}\left[\left\| \frac{1}{M} \sum_{i \in \mathcal{M}_t} \boldsymbol{\Delta}_{t-\tau_t^i}^i \right\|^2 \right] \\
&\leq \frac{2K\eta_l^2}{M} \sigma^2 + \frac{2\eta_l^2 (N-M)}{NM(N-1)} \Bigg[15NK^3 L^2 \eta_l^2 (\sigma^2 + 6K\sigma_g^2) + (90NK^4 L^2 \eta_l^2 + 3K^2) \\
&\quad \cdot \sum_{i=1}^{N} \mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] + 3NK^2 \sigma_g^2 \Bigg] + \frac{2\eta_l^2 (M-1)}{NM(N-1)} \mathbb{E}\left[\left\| \sum_{i=1}^{N} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i) \right\|^2 \right].
\end{aligned}$$

Proof. The proof of Lemma C.3 is similar to the proof of Lemma C.6 in (Wang et al., 2022b). □

Lemma C.5. *(This lemma follows from Lemma 3 in FedAdam (Reddi et al., 2021). For local learning rate which satisfying* $\eta_l \leq \frac{1}{8KL}$, *the local model difference after* k *(*$\forall k \in \{0, 1, ..., K-1\}$*) steps local updates satisfies*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{x}_{t,k}^i - \boldsymbol{x}_t\|^2] \leq 5K\eta_l^2 (\sigma_l^2 + 6K\sigma_g^2) + 30K^2 \eta_l^2 \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]. \tag{75}$$

Proof. The proof of Lemma C.5 is similar to the proof of Lemma 3 in Reddi et al. (2021). □

Lemma C.6. *If assuming that the clients' participation distributions are simulated as independently uniform distribution, then the sequence* $\boldsymbol{G}_s = \sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})$ *has the following upper bound,*

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&\le \frac{M(N-M)}{N-1}\cdot\Big[15K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]\Big]\\
&\quad+\frac{3M(N-M)}{N-1}K^2\sigma_g^2+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big].
\end{aligned}$$

Proof. We begin with the proof similar to the partial participation with sampling without replacement,

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{j\in\mathcal{M}_s}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&=\frac{M(N-M)}{N(N-1)}\sum_{j=1}^{N}\mathbb{E}\Big[\Big\|\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&\le\frac{M(N-M)}{N(N-1)}\Big[15NK^3\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]+3NK^2\sigma_g^2\Big]\\
&\quad+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&\le\frac{M(N-M)}{N(N-1)}\Big[15NK^3\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]+3NK^2\sigma_g^2\Big]\\
&\quad+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&\le\frac{M(N-M)}{N(N-1)}\Big[15NK^3\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]+3NK^2\sigma_g^2\Big]\\
&\quad+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&\le\frac{M(N-M)}{N-1}\Big[15K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]+3K^2\sigma_g^2\Big]\\
&\quad+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big]\\
&=\frac{M(N-M)}{N-1}\Big[15K^3L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+(90K^4L^2\eta_l^2+3K^2)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{s-\tau_s^j})\|^2]\Big]\\
&\quad+\frac{3M(N-M)}{N-1}K^2\sigma_g^2+\frac{M(M-1)}{N(N-1)}\mathbb{E}\Big[\Big\|\sum_{j=1}^{N}\sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}^j_{s-\tau_s^j,k})\Big\|^2\Big].
\end{aligned}\tag{76}$$

□

D. Additional Experiments

Tables 8 and 9 present results computed over experiments with 3 different random seeds. Tables 8 and 9 compare the performance of various federated learning methods, on the test accuracy of the ResNet-18 model across CIFAR-10 and CIFAR-100 datasets with heterogeneous data distributions. It is observed that the delay-adaptive FADAS (abbreviated as delay-adaptive FADAS) consistently outperforms the other methods. The consistency of FADAS performance under *large worst-case* delay settings indicates its reliability and potential for practical applications in federated learning environments with diverse and asynchronous model updates.

D.1. Additional Results

Table 8. The test accuracy on training ResNet-18 model on CIFAR-10 dataset with two data heterogeneity levels in a large worst-case delay scenario for 500 communication rounds. We abbreviate delay-adaptive FADAS to FADAS_{da} in this and subsequent tables. We conduct experiments on three seeds, and we report the average accuracy and standard derivation.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	50.29 ± 6.86	59.75 ± 13.40
FedBuff	44.92 ± 5.26	46.94 ± 0.99
FADAS	70.57 ± 2.04	75.97 ± 2.64
FADAS_{da}	$\mathbf{72.64} \pm 1.00$	$\mathbf{80.26} \pm 0.68$

Table 9. The test accuracy on training ResNet-18 model on CIFAR-100 dataset with two data heterogeneity levels in a *large worst-case* delay scenario for 500 communication rounds. We conduct experiments on three seeds, and we report the average accuracy and standard derivation.

Method	Dir(0.1) Acc. & std.	Dir (0.3) Acc. & std.
FedAsync	46.25 ± 4.33	43.22 ± 10.75
FedBuff	15.97 ± 2.44	28.58 ± 4.74
FADAS	47.85 ± 0.69	52.80 ± 1.15
FADAS_{da}	$\mathbf{51.55} \pm 1.03$	$\mathbf{56.01} \pm 0.95$

D.2. Implementation Details

Details of applying adaptive learning rate. During our experiments, we found that choosing a relatively small global learning rate η yields better results for adaptive FL methods (hyper-parameter details can be found in the following). To scale the learning rate down for the model update with larger delays, we directly scale down the learning rate for this step to $\eta/\tau_t^{\max}$, which is shown in (77),

$$\eta_t = \begin{cases} \eta & \text{if } \tau_t^{\max} \leq \tau_c, \\ \min\left\{\eta, \frac{\eta}{\tau_t^{\max}}\right\} & \text{if } \tau_t^{\max} > \tau_c. \end{cases} \tag{77}$$

Hyper-parameter Settings. We conduct detailed hyper-parameter searches to find the best hyper-parameter for each baseline. We grid the local learning rate η_l from $\{0.001, 0.003, 0.01, 0.03, 0.1\}$, and global learning rate $\eta = 1$ for SGD-based method. We grid the local learning rate η_l from $\{0.003, 0.01, 0.03, 0.1\}$ and global learning rate η from $\{0.0001, 0.0003, 0.001, 0.003\}$ for adaptive method. For the global adaptive optimizer, we set $\beta_1 = 0.9$, $\beta_1 = 0.99$, and we set $\epsilon = 10^{-8}$. Table 10 summarizes the hyper-parameter details in our experiments.

Table 10. Hyper-parameters details for vision tasks.

CIFAR-10 (mild delay)								
	FedAsync		FedBuff		FADAS		FADAS_{da}	
Models & Dir(α)	η_l	η	η_l	η	η_l	η	η_l	η
ResNet-18 & Dir(0.1)	0.003	1	0.03	1	0.1	0.0003	0.1	0.001
ResNet-18 & Dir(0.3)	0.01	1	0.03	1	0.1	0.0003	0.1	0.001
CIFAR-100 (mild delay)								
	FedAsync		FedBuff		FADAS		FADAS_{da}	
Models & Dir(α)	η_l	η	η_l	η	η_l	η	η_l	η
ResNet-18 & Dir(0.1)	0.01	1	0.03	1	0.1	0.0003	0.1	0.001
ResNet-18 & Dir(0.3)	0.01	1	0.03	1	0.1	0.0003	0.1	0.001
CIFAR-10 (large worst-case delay)								
	FedAsync		FedBuff		FADAS		FADAS_{da}	
Models & Dir(α)	η_l	η	η_l	η	η_l	η	η_l	η
ResNet-18 & Dir(0.1)	0.003	1	0.03	1	0.1	0.0001	0.1	0.001
ResNet-18 & Dir(0.3)	0.003	1	0.03	1	0.1	0.0001	0.1	0.001
CIFAR-100 (large worst-case delay)								
	FedAsync		FedBuff		FADAS		FADAS_{da}	
Models & Dir(α)	η_l	η	η_l	η	η_l	η	η_l	η
ResNet-18 & Dir(0.1)	0.003	1	0.03	1	0.1	0.0001	0.1	0.001
ResNet-18 & Dir(0.3)	0.001	1	0.03	1	0.1	0.0001	0.1	0.001

Table 11. Hyper-parameters details for language tasks.

	FedAsync		FedBuff		FADAS		FADAS_{da}	
Datasets	η_l	η	η_l	η	η_l	η	η_l	η
RTE	0.01	1	0.01	1	0.01	0.005	0.01	0.01
MRPC	0.001	1	0.01	1	0.01	0.001	0.01	0.002
SST-2	0.001	1	0.001	1	0.1	0.0005	0.1	0.001