# Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration

**Yujia Wang** [1] **Yuanpu Cao** [1] **Jingcheng Wu** [2] **Ruoyu Chen** [2] **Jinghui Chen** [1]

## Abstract

Asynchronous federated learning, which enables local clients to send their model update asynchronously to the server without waiting for others, has recently emerged for its improved efficiency and scalability over traditional synchronized federated learning. In this paper, we study how the asynchronous delay affects the convergence of asynchronous federated learning under non-i.i.d. distributed data across clients. We first analyze the convergence of a general asynchronous federated learning framework under a practical nonconvex stochastic optimization setting. Our result suggests that the asynchronous delay can largely slow down the convergence, especially when the data heterogeneity is high. To further improve the convergence of asynchronous federated learning with heterogeneous data distribution, we then propose a novel asynchronous federated learning method with a cached update calibration. Particularly, we let the server cache the latest update for each client and reuse these variables for calibrating the global update at each round. We theoretically prove the convergence acceleration for our proposed method under nonconvex stochastic settings and empirically demonstrate its superior performances compared to standard asynchronous federated learning. Moreover, we also extend our method with a memory-friendly adaption in which the server only maintains a quantized cached update for each client for reducing the server storage overhead.

## 1. Introduction

Federated Learning (McMahan et al., 2017) has become an increasingly popular large-scale machine learning paradigm where machine learning models are trained on multiple edge clients guided by a central server. FedAvg (McMahan et al., 2017), also known as Local SGD (Stich, 2018), is one of the most popular federated optimization methods, where each client locally performs multiple steps of SGD updates followed by the synchronous server aggregation of the local models. However, the traditional synchronous aggregation scheme may cause efficiency and scalability issues as the server need to wait for all participating clients to complete the task before conducting the global update step. This promotes the development of asynchronous federated learning methods such as FedAsync (Xie et al., 2019), and FedBuff (Nguyen et al., 2022), which adopt flexible aggregation schemes and allow clients to asynchronously send back their model update and thus improve the overall training efficiency and scalability.

Such an asynchronous aggregation scheme does not come with no costs: the asynchronous delay, which describes the fact that the delayed local model update could be computed based on a past global model rather than the current global model, slows down the convergence of asynchronous federated learning. Moreover, the negative impact of the asynchronous delay on the convergence gets even worse when the training data are non-i.i.d. distributed across clients. This is intuitive since empirical observation suggests that the global model changes more significantly in adjacent rounds when the data heterogeneity is high. Consequently, the asynchronous delay would cause the delayed local model update to be more outdated and inconsistent with the current global model, hence worsening the overall model convergence. Therefore, it is crucial to tackle the data heterogeneity issue in asynchronous federated learning, not only for reducing the negative impact of data heterogeneity itself but also for reducing the impact of the asynchronous delay and improving the overall convergence.

In this work, we rigorously study how the asynchronous delay affects the convergence of asynchronous federated learning under non-i.i.d. distributed data across clients. We first conduct the theoretical analysis of a general asynchronous federated learning framework under a nonconvex stochastic setting and verify that the effect brought by asynchronous delay would be amplified by the highly non-i.i.d. distributed

[1]The Pennsylvania State University [2]Carnegie Mellon University. Correspondence to: Jinghui Chen <jzc5917@psu.edu>.

*Workshop of Federated Learning and Analytics in Practice, colocated with $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. 

data. Inspired by the incremental gradient in SAGA (Defazio et al., 2014), we then develop a novel asynchronous federated learning method, Cache-Aided Asynchronous Federated Learning (CA$^2$FL), for improving the convergence degradation caused by the joint effect of data heterogeneity and asynchronous delay. In CA$^2$FL, the server maintains the latest update from each client and reuses this cached update for calibrating the global update. The proposed CA$^2$FL does not change the local training steps on clients and only modifies the global aggregation steps. Therefore, the proposed CA$^2$FL does not incur extra communication and computation overhead on clients, and it does not raise additional privacy concerns than the traditional synchronous and asynchronous federated learning methods. Moreover, we extend our CA$^2$FL to a memory-friendly adaption for further scalability improvement. We summarize our contribution in this paper as follows:

- We investigate the convergence property of the general asynchronous federated learning framework, which benefits from the flexible aggregation scheme with improved efficiency and scalability, under non-i.i.d. distributed data across clients. We demonstrate that the asynchronous delay can theoretically slow down the convergence and such an impact could be further amplified by the highly non-i.i.d. distributed data.
- To tackle the convergence degradation caused by the joint effect of data heterogeneity and asynchronous delay, we propose a novel asynchronous federated aggregation method with cached update calibrations (CA$^2$FL) in which the server maintains cache updates for each client and reuse the cached update for global aggregation calibration. We prove that with the help of cached updates, our proposed method can significantly improve the convergence rate under nonconvex stochastic settings. Empirical experiments on benchmark datasets and models verify the effectiveness of the proposed method.
- We extend our proposed CA$^2$FL to a memory-friendly cached update calibration method, MF-CA$^2$FL, where the server only maintains the quantized cached update instead of the full-size one. We show that MF-CA$^2$FL can achieve very similar performance and final accuracy as CA$^2$FL, with much fewer memory costs.

## 2. Preliminaries

**Federated learning framework.** In general federated learning framework, we aim to minimize the following objective through $N$ local clients:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\xi\sim\mathcal{D}_i}[F_i(\boldsymbol{x};\xi_i)], \quad (2.1)$$

where $\boldsymbol{x}$ represents the model parameters with $d$ dimensions, $F_i(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[F_i(\boldsymbol{x},\xi_i)]$ represents the local loss function corresponding to client $i$ and let $\mathcal{D}_i$ denotes the local data distribution on client $i$. In this work, we focus on the non-convex optimization problem with heterogeneous data distributions, i.e., $F_i$ are non-convex and the local data distributions $\mathcal{D}_i$ and $\mathcal{D}_j$ are non-i.i.d. distributed for different client $i$ and $j$. FedAvg (McMahan et al., 2017) is a popular synchronous optimization algorithm to solve Eq. 2.1, where each participating client performs local SGD updates, and the server performs global averaging steps after receiving all the updates from assigned clients.

**General asynchronous federated learning framework.** Asynchronous federated learning has been introduced to facilitate efficiency and scalability for clients in solving Eq. 2.1 asynchronously. In asynchronous federated learning, clients are allowed to train and synchronize local models at their own pace. Several works such as FedAsync (Xie et al., 2019) and FedBuff (Nguyen et al., 2022) have explored different aspects of asynchronous federated learning. Specifically, FedAsync (Xie et al., 2019) studied an algorithm that the server would immediately update the global model once it receives a local model from an arbitrary client while aggregating individual client updates may cause some privacy issues. FedBuff (Nguyen et al., 2022) studied an asynchronous federated learning method with differential privacy and secure aggregation consideration, thus we generalize FedBuff (without differential privacy) into this framework. We summarize a general asynchronous federated learning framework in Appendix B, which is structured by enabling the server to collect several clients' updates for updating one step of a global model.

**Heterogeneous across clients.** Several works (Karimireddy et al., 2020b;a; Acar et al., 2021; Wang et al., 2020b) have shown that synchronized federated learning methods suffer from convergence and empirical degradation when data is heterogeneously distributed across local clients. In particular, the model variation may be more significant when only a subset of clients contribute to a round of global updates. This issue of model inconsistency also occurs in asynchronous federated learning and may even become worse with the existing of gradient delay $\tau_t^i$, since the model used for local gradient computation is usually different from the current global model, which makes local updates less representative of the global update direction. This intuition has also been studied in the convergence rate under stochastic non-convex settings for general asynchronous federated learning as informally stated below.

**Theorem 2.1** (*Informal*, formal statement and proof in Appendix B). *Assume that* $\forall i \in [N]$, $F_i$ *is smooth under a common assumption. Let* $\sigma^2$ *and* $\sigma_g^2$ *be the stochastic and global variance, and let* $\tau_{\max} = \max_{t\in[T], i\in\mathcal{S}_t}\{\tau_t^i\} < \infty$ *be the maximum gradient delay, define* $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ *and* $f_0 = f(\boldsymbol{x}_1)$ *. If picking the local learning rate* $\eta = \Theta(\sqrt{KM})$ *and*

$\eta_l = \Theta(1/\sqrt{T}K)$, *then the global rounds of Algorithm 2 satisfy* $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] = \mathcal{O}\big(\frac{[(f_0-f_*)+\sigma^2]}{\sqrt{TKM}}\big) + \mathcal{O}\big(\frac{\sigma^2+K\sigma_g^2}{TK}\big) + \mathcal{O}\big(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2\big) + \mathcal{O}\big(\frac{K\tau_{\max}^2\sigma_g^2+\tau_{\max}^2\sigma^2}{T}\big)$.

*Remark* 2.2. Theorem 2.1 presents the informal convergence analysis for Algorithm 2 w.r.t. global communication round $T$, local steps $K$ and the update accumulation amount $M$. From Theorem 2.1, it can be seen that the maximum delay $\tau_{\max}$ term indeed affects the overall convergence of the asynchronous federated learning algorithm. Particularly, the last term involves joint effect term $\mathcal{O}(K\tau_{\max}^2\sigma_g^2/T)$ where the global variance $\sigma_g^2$ and the maximum delay $\tau_{\max}$ are multiplied together [1]. This implies that the convergence degradation brought by the asynchronous delay $\tau_{\max}$ is amplified by the high data heterogeneity (large $\sigma_g$). If data are i.i.d. distributed across clients, i.e., $\sigma_g = 0$, then $\mathcal{O}(K\tau_{\max}^2\sigma_g^2/T)$ term vanishes to 0. On the other hand, if data are non-i.i.d. distributed, i.e., $\sigma_g \neq 0$, the term $\mathcal{O}(K\tau_{\max}^2\sigma_g^2/T)$ will largely slow down the overall convergence (in fact, when $T \leq KM$, this term would become the dominant term in the convergence rate). This verifies our intuition that the data heterogeneity can worsen the impact of asynchronous delay and jointly deteriorate the convergence, which motivates us to develop a novel method for reducing such joint effects and improving the convergence for asynchronous federated learning.

## 3. Proposed Method: CA$^2$FL

To address the challenges of data heterogeneity and gradient delay across clients and achieve better convergence in asynchronous federated learning, we propose a novel Cache-Aided Asynchronous FL (CA$^2$FL) method. The proposed CA$^2$FL enables the server to maintain and reuse the cached updates for global update calibration. Algorithm 1 summarizes our proposed CA$^2$FL. In general, the CA$^2$FL largely follows the Asynchronous FL framework in Algorithm 2, while the main difference between our proposed CA$^2$FL and Algorithm 2 lies primarily in the global update steps. Specifically, we introduce a *global calibration* process in Line 9 and incorporate steps for *cached variable updating* shown in Line 11-12.

**Global calibration.** In CA$^2$FL, the server maintains a latest cached update for each client, and reuses this cached update as an approximation of each client's contribution to the current round's update. Specifically, at global round $t$, denote $\boldsymbol{h}_t^i$ as the latest cached variable for client $i$ and $\boldsymbol{h}_t$ as the global cached variable which is the average of $\boldsymbol{h}_t^i$ among all clients, i.e., $\boldsymbol{h}_t = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_t^i$, let $\mathcal{S}_t$ represent a set of clients in which the server received the update at round $t$. The server updates the global model to $\boldsymbol{x}_{t+1}$ by using a calibrated variable $\boldsymbol{v}_t$, which is a linear combination in terms of the global cached variable $\boldsymbol{h}_t$ and the latest received model update difference $\Delta_{t-\tau_t^i}^i$ and $\boldsymbol{h}_t^i$.

**Cached variable update.** The server then updates $\boldsymbol{h}_t^i$ based on whether received the update from client $i$ or not (Line 11 in Algorithm 1): if the server received $\Delta_{t-\tau_t^i}^i$ from client $i$, then the server updates the cached variable for it, i.e., $\boldsymbol{h}_{t+1}^i = \Delta_{t-\tau_t^i}^i$, otherwise the server keeps the state variable unchanged as $\boldsymbol{h}_{t+1}^i = \boldsymbol{h}_t^i$. This update rule for cached variable enforces the server maintains the latest $\Delta_{t-\tau_t^i}^i$ for each client for global update calibration.

**Algorithm 1** Cached-Aided Asynchronous FL

**Input:** initial point $\boldsymbol{x}_1$, local step size $\eta_l$, global stepsize $\eta$, server concurrency $M_c$, server updates after receive $M$ updates from clients

1: Initialize sampled set with $|\mathcal{M}_1| = M_c$ clients, send server initial model $\boldsymbol{x}_1$ to active clients
2: **repeat**
3: Initialize $\mathcal{S}_t = \emptyset$, clients in $\mathcal{M}_t$ perform local SGD updates based on Algorithm 3
4: **if** receive client update **then**
5: Server receive client update $\Delta_{t-\tau_t^i}^{i_t}$ from client $i_t$: $\Delta_t \leftarrow \Delta_t + \Delta_{t-\tau_t^i}^{i_t}$
6: $m \leftarrow m+1$, $\mathcal{S}_t \leftarrow \mathcal{S}_t \cup \{i_t\}$
7: **end if**
8: **if** $m = M$ **then**
9: Update $\boldsymbol{v}_t = \boldsymbol{h}_t + \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}(\Delta_{t-\tau_t^i}^i - \boldsymbol{h}_t^i)$, where $\boldsymbol{h}_t = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_t^i$
10: Update global model $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta\boldsymbol{v}_t$
11: Update clients' cached variables: for $j \notin \mathcal{S}_t$, $\boldsymbol{h}_{t+1}^j = \boldsymbol{h}_t^j$, for $i \in \mathcal{S}_t$, $\boldsymbol{h}_{t+1}^i = \Delta_{t-\tau_t^i}^{i_t}$
12: Reset $m \leftarrow 0$, $t \leftarrow t+1$
13: Sample client $\mathcal{S}_{t+1} \subseteq [N]/\mathcal{M}_t \cup \mathcal{S}_t$, update $\mathcal{M}_{t+1} \leftarrow \mathcal{M}_t/\mathcal{S}_t \cup \mathcal{S}_{t+1}$, and broadcast global model $\boldsymbol{x}_t$ to clients in $\mathcal{S}_{t+1}$
14: **end if**
15: **until** Convergence

**Discussion.** The design for the calibration and cached variables felt somewhat similar to SAGA (Defazio et al., 2014), a well-recognized stochastic variance-reduction method that stores previously computed gradients and leverages them for reducing the gradient variance.

Our design looks like a special form of SAGA by treating model update difference $\Delta_{t-\tau_t^i}^i$ as gradients and applied globally over different clients. However, it is important to note that our method does not adhere to the properties

[1] It is worth noting that our dependency of $\tau_{\max}$ is on the same order as FedBuff (Nguyen et al., 2022), and we can further obtain a linear dependency of $\tau_{\max}$ as in Koloskova et al. (2022) with adapting on the learning rate.

of unbiased incremental gradients that SAGA mainly relies on for its variance reduction purposes, which makes our theoretical analysis non-trivial and different from that of SAGA. Therefore, $\text{CA}^2\text{FL}$ should not be considered as a direct application of SAGA to asynchronous federated learning.

Note that $\text{CA}^2\text{FL}$ does not require extra communication and computation overhead on clients, and it is compatible with privacy persevering approaches such as differential privacy and secure aggregation. However, one drawback is that $\text{CA}^2\text{FL}$ introduces extra memory overhead on the server since it needs to store a cached update for each client. To reduce this storage overhead, we further extend the proposed $\text{CA}^2\text{FL}$ to a memory friendly adaption in Appendix C, which demonstrates that $\text{CA}^2\text{FL}$ can maintain overall good performance without the need to maintain full-size cached updates.

## 4. Convergence Analysis

Due to space limitations, we will introduce the assumptions needed for the convergence analysis in Appendix.

**Theorem 4.1** (*Informal* Convergence analysis for Algorithm 1)**.** *Assume that $\forall i \in [N]$, $F_i$ is smooth under a common assumption. Let $\sigma^2$ and $\sigma_g^2$ be the stochastic and global variance, let $\tau_{\max} = \max_{t\in[T], i\in\mathcal{S}_t}\{\tau_t^i\} < \infty$ be the maximum gradient delay and let $\zeta_{\max} = \max_{t\in[T], i\in\mathcal{S}_t}\{\zeta_t^i\} < \infty$ represents the maximum state delay. If the local learning rate $\eta = \Theta(\sqrt{KM})$ and $\eta_l = \Theta(1/\sqrt{T}K)$ then the global rounds of Algorithm 2 satisfy*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&= \mathcal{O}\bigg(\frac{f_0 - f_*}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2 + K\sigma_g^2}{TK}\bigg) \\
&\quad + \mathcal{O}\bigg(\frac{\tau_{\max}^2\sigma^2}{T}\bigg) + \mathcal{O}\bigg(\frac{\zeta_{\max}^2(N-M)^2\sigma^2}{TN^2}\bigg), \qquad (4.1)
\end{aligned}
$$

*where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$.*

*Remark* 4.2. Theorem 4.1 suggests that with a sufficient amount of global communication rounds $T$, i.e., $T \geq KM$, our proposed $\text{CA}^2\text{FL}$ method achieves a desired convergence rate of $\mathcal{O}(\frac{1}{\sqrt{TKM}})$ w.r.t. global communication round $T$, local steps $K$ and the update accumulation amount $M$, which matches the convergence rate in traditional synchronous federated learning baselines (Yang et al., 2021; Reddi et al., 2021; Jhunjhunwala et al., 2022).

*Remark* 4.3. Compared with Theorem 2.1, we notice that in Theorem 4.1, the joint effect term $\mathcal{O}(K\tau_{\max}^2\sigma_g^2/T)$ no longer exists, while the asynchronous delay $\tau_{\max}$ only relates to the stochastic noise $\sigma$. This suggests that our proposed $\text{CA}^2\text{FL}$ can benefit from the design of reusing the cached update for global update calibration, which tackles the data heterogeneity issue across clients and reduces the joint impact caused by the asynchronous delay and data heterogeneity. Note that our design also contributes to the general data heterogeneity issue in that the $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2)$ term in Theroem 4.1 also gets smaller. Together, those two improvements finally lead to a better convergence rate for our proposed $\text{CA}^2\text{FL}$ algorithm.

## 5. Memory Friendly Cached-Aided Asynchronous FL

While $\text{CA}^2\text{FL}$ successfully tackles the data heterogeneity issue in Asynchronous FL, it involves extra memory costs for maintaining the cached variable for each client on the server. However, this memory overhead can pose challenges when applying $\text{CA}^2\text{FL}$ in practice, especially for large models with massive trainable parameters. To overcome this memory overhead, we extend the proposed $\text{CA}^2\text{FL}$ to a memory-friendly adaption method (MF-$\text{CA}^2\text{FL}$). The main difference between $\text{CA}^2\text{FL}$ and MF-$\text{CA}^2\text{FL}$ lies in whether the server maintains a full-size or a quantized latest update. Specifically, in MF-$\text{CA}^2\text{FL}$, after the client $i$ obtains the model differences $\Delta_{t-\tau_t^i}^i$ and sends it to the server, the server quantizes $\Delta_{t-\tau_t^i}^i$ to $\mathcal{Q}(\Delta_{t-\tau_t^i}^i)$ via unbiased quantization approaches [2] and keeps $\mathcal{Q}(\Delta_{t-\tau_t^i}^i)$ in memory. The server updates the global calibration variable $\boldsymbol{v}_t$ same as $\text{CA}^2\text{FL}$. Note that for each global round $t$, the server updates the quantized $\mathcal{Q}(\Delta_{t-\tau_t^i}^i)$ as the cached update, i.e., $\boldsymbol{h}_{t+1}^i = \mathcal{Q}(\Delta_{t-\tau_t^i}^i), \forall i \in \mathcal{S}_t$, that being said, the cached variable $\boldsymbol{h}_{t+1}^i$ for each client represents the latest quantized model update difference. Therefore, compared to $\text{CA}^2\text{FL}$, this memory-friendly adaption effectively reduces the memory overhead. Due to space limits, we summarize the detailed MF-$\text{CA}^2\text{FL}$ algorithm and provide a complete theoretical analysis for convergence guarantee in Appendix C.

## 6. Experimental Results

**Datasets, models, and methods.** We present the experimental results on the CIFAR-10 (Krizhevsky et al., 2009) dataset where we evaluate experiments on non-i.i.d. data distributions by a Dirichlet distribution partitioned strategy with parameter $\alpha = 0.3$ similar to Wang et al. (2020a;b). We adopt the same CNN network as in Wang & Ji (2022) and ResNet-18 network (He et al., 2016). We compare our proposed $\text{CA}^2\text{FL}$ and MF-$\text{CA}^2\text{FL}$ with the general Asynchronous federated learning baseline (Algorithm 2). Note that this asynchronous FL baseline is essentially the same as

[2] Due to space limits, we leave detailed discussion of the quantization approaches in Appendix C.

FedBuff without differential privacy (Nguyen et al., 2022) when limiting the concurrency of clients. Due to the space limit, we leave additional experiments on more datasets and models together with the experiment details in Appendix D.

**Main Results.** Table 1 shows the overall performance on training CIFAR-10 with a CNN model and the ResNet-18 model. We observe that the proposed CA$^2$FL shows improvement upon the general Asynchronous FL baseline, and the proposed MF-CA$^2$FL with 8 bits and 4 bits quantization maintains the superior performance of the cached update with just 0.1%-0.3% loss decreasing comparing to the proposed CA$^2$FL method, while reduces the memory overhead by up to 8 times compared to CA$^2$FL. This demonstrates that our proposed CA$^2$FL and MF-CA$^2$FL with 8 bits or 4 bits quantization achieve overall better performance than the general asynchronous federated learning method.

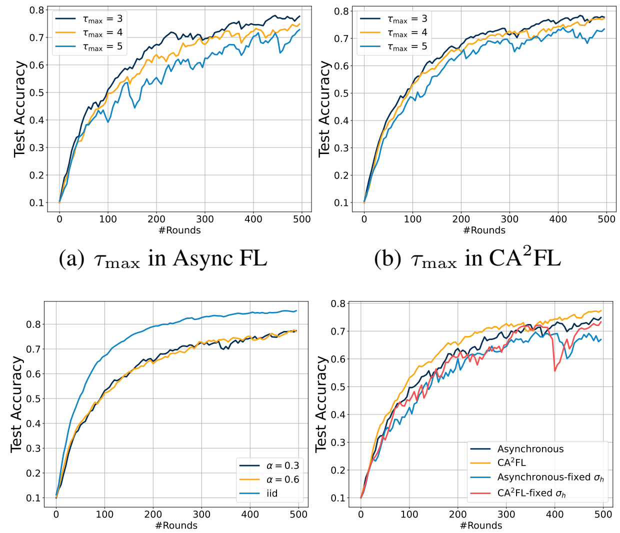
*Table 1.* The test accuracy of training CNN and ResNet-18 models on CIFAR-10. We report the final accuracy of training 500 global rounds, and the global round when achieves the desired accuracy.

| Method & Model | CNN | | ResNet-18 | |
|---|---|---|---|---|
| | Acc. | R#(50%) | Acc. | R# (80%) |
| Asynchronous FL | 50.23 | **284** | 74.22 | 500+ |
| CA$^2$FL | **53.66** | 294 | **77.16** | **449** |
| MF-CA$^2$FL (8 bits) | 53.54 | 314 | 75.09 | 461 |
| MF-CA$^2$FL (4 bits) | 53.38 | 329 | 74.18 | 500+ |
| MF-CA$^2$FL (2 bits) | 41.55 | 500+ | 51.19 | 500+ |

**Ablation Studies.** We conduct ablation studies to investigate the effect of maximum asynchronous delay $\tau_{\max}$, the effect of data heterogeneity, and the effect of different delay sampling strategies. Due to constraints on space, we provide detailed ablation results and discussions in Appendix D. Figure 1 shows curves of test accuracy for several ablations studies. From plots (a) and (b) we can observe that compared to the general Asynchronous federated learning method, our proposed is less sensitive to the variation of maximum delay [3]. This suggests that the delay could have a relatively weaker impact on the overall model performance in CA$^2$FL. We show the different levels of data heterogeneity in the plot (c), Figure 3. For plot (d), we investigate the impact of letting all clients' wall-time delay sampled from the same distribution $\sigma_h \sim$ halfnorm(5) or letting clients' wall-time delay randomly sampled from different half normal distributions. We observe that sampling from the same distributions worsens the overall performance of both our proposed CA$^2$FL and the general Asynchronous federated learning baseline [4].

[3]Due to the algorithm structure, we cannot directly control the maximum delay, instead, we adjust the overall concurrency $M_c$ and report the result with fixed model accumulation amount $M$ and different levels of concurrency $M_c$.

[4]We further discuss this in Appendix D.




(a) $\tau_{\max}$ in Async FL (b) $\tau_{\max}$ in CA$^2$FL

(c) Data heterogeneity (d) Delay simulations

*Figure 1.* Ablation study with several context: (a) the effect of maximum delay in the Asynchronous FL baseline; (b) the effect of maximum delay in the proposed CA$^2$FL; (c) the effect of data heterogeneity in CA$^2$FL; (d) the effect of different delay $\sigma_h$ sampling methods.

## 7. Conclusions

In this paper, we first investigate the convergence of general asynchronous federated learning under heterogeneous data distributions and we show that the data heterogeneity amplifies the negative impact of asynchronous delay which slows down the convergence of asynchronous federated learning. To address this convergence degradation issue, we propose a novel asynchronous federated learning method, CA$^2$FL, which involves caching and reusing previous updates for global calibration. We provide theoretical analysis under non-convex stochastic settings that demonstrate the significant convergence improvement of our proposed CA$^2$FL. Moreover, we extend the proposed CA$^2$FL to a memory-friendly adaption, MF-CA$^2$FL, for reducing the storage overhead caused by caching the latest update. Empirical results demonstrate the superior performance of the proposed CA$^2$FL compared to general asynchronous federated learning, and it also shows that the proposed MF-CA$^2$FL could largely save the memory overhead while maintaining the superior performance benefits from the cached update.

## 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.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

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.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., and Xu, C.-Z. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10112–10121, 2022.

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.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

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.

Jhunjhunwala, D., Sharma, P., Nagarkatti, A., and Joshi, G. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pp. 906–916. PMLR, 2022.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

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, 2020b.

Khanduri, P., Sharma, P., Yang, H., Hong, M., Liu, J., Rajawat, K., and Varshney, P. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34: 6050–6061, 2021.

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.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

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. URL http://jmlr.org/papers/v19/17-650.html.

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

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*, 2019.

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.

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.

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

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., 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 unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19124–19137, 2022.

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, 2022.

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. URL https://openreview.net/forum?id=jDdzh5ul-d.

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

Yang, Z., Chen, M., Saad, W., Hong, C. S., and Shikh-Bahaei, M. Energy efficient federated learning over wireless communication networks. *IEEE Transactions on Wireless Communications*, 20(3):1935–1949, 2020.

# A. Related Work and Preliminaries

**Synchronous FL.** Federated learning (Konečný et al., 2016) play a critical role in collaboratively training models at edge devices with potential privacy protections. Basic optimization methods for federated learning include SGD-based global optimizer, e.g., FedAvg (McMahan et al., 2017) (a.k.a. Local SGD (Stich, 2018) and its variants (Li et al., 2019; Yang et al., 2021), adaptive gradient optimization based global optimizer such as FedAdam (Reddi et al., 2021), FedAGM (Tong et al., 2020) and FedAMS (Wang et al., 2022). Recently, several works address the data heterogeneity issue through several aspects. For example, FedProx (Li et al., 2020) adds a proximate term to align the local model with the global one, and FedDyn (Acar et al., 2021) involves dynamic regularization term for local and global model consistency. FedNova (Wang et al., 2020b) proposes a normalized averaging mechanism that reduces objective inconsistency with heterogeneous data. Moreover, several works study to eliminate the client drift caused by data heterogeneity from the aspect of variance reduction such as Karimireddy et al. (2020b;a); Khanduri et al. (2021); Cutkosky & Orabona (2019); Jhunjhunwala et al. (2022). They introduce additional control variables to track and correct the local model shift during local training, but they require extra communication costs for synchronizing these control variables. Besides, FedDC (Gao et al., 2022) involves both dynamic regularization terms and local drift variables for model correction.

**Asynchronous SGD and Asynchronous FL.** Asynchronous optimization methods such as asynchronous SGD and its variants have been discussed for many years. For example, Hogwild! SGD (Niu et al., 2011) studies a coordinate-wise asynchronous method without any locking which allows processors access to shared memory and provides the possibility of overwriting each other's work, and Nguyen et al. (2018) provided a tight convergence analysis for SGD and Hogwild! algorithm. Some other works focusing on the theoretical analysis for the asynchronous SGD such as Mania et al. (2017); Stich et al. (2021). Leblond et al. (2018) studies the SAGA method in the context of asynchronous SGD and demonstrates the theoretical convergence improvement of the asynchronous SAGA. Glasgow & Wootters (2022) explored SAGA methods in the context of asynchronous distributed-data settings provided a theoretical analysis under (strongly) convex loss functions. In the context of federated learning, the system heterogeneity across clients, e.g., the computation capabilities and communication bandwidths, limits the efficiency and practicality. Hence the asynchronous federated learning aggregation methods have been raised for adjusting for the flexibility and scalability consideration. For example, FedAsync (Xie et al., 2019) is proposed for clients to update asynchronously to the server, and FedBuff (Nguyen et al., 2022) is extended to a buffered asynchronous aggregation strategy. Anarchic Federated Averaging (AFA) (Yang et al., 2022) is another related work focusing on letting the clients decide when and whether to participate in global training. Moreover, there are several works studying the theoretical convergence analysis in asynchronous federated learning with arbitrary delay (Avdiukhin & Kasiviswanathan, 2021; Mishchenko et al., 2022) or the complete theoretical analysis under various assumptions (Koloskova et al., 2022).

**Preliminaries.** As we mentioned in Section 2, here we summarize the general asynchronous federated learning methods in Algorithm 2. In Algorithm 2, the server initializes by selecting an active client set $\mathcal{M}_1$ with concurrency $M_c$ [5] and assigning the initial model $\boldsymbol{x}_1$ to these clients. Throughout the algorithm, all clients conduct $K$ steps of local training asynchronously (Algorithm 3) at their own pace. This means each client trains the local model with the previously assigned global model $\boldsymbol{x}_{t-\tau_t^i}$, where $\tau_t^i$ represents the gradient delay, i.e., the difference between the round when client $i$ start to compute the gradient and the round that the update difference $\Delta_{t-\tau_t^i}^i$ from client $i$ is received by the server. The server does not immediately update the global model once receiving a client's update, instead, it accumulates the model update difference $\Delta_t$ (Line 5 in Algorithm 2) and updates the global model $\boldsymbol{x}_{t+1}$ once it receives $M$ updates from clients in $\mathcal{S}_t$ (Lines 10-12 in Algorithm 2). After that, a subset of clients $\mathcal{S}_{t+1}$ are assigned with the latest update $\boldsymbol{x}_{t+1}$, as shown in Line 11 of Algorithm 2. Note that this client assignment ensures that a client can only be sampled once in a specific global round $t$. Once it updates the update difference to the server, it becomes eligible for immediate assignment in the subsequent round of training.

# B. Convergence Analysis for Asynchronous FL and CA$^2$FL

First, we state several general assumptions needed for the convergence analysis.

**Assumption B.1** (Smoothness). Each loss function on the $i$-th worker $F_i(\boldsymbol{x})$ is $L$-smooth, i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\left|F_i(\boldsymbol{x}) - F_i(\boldsymbol{y}) - \langle \nabla F_i(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle\right| \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

[5] The concurrency implies that the maximum simultaneously active clients is $M_c$.

**Algorithm 2** Asynchronous FL

**Input:** initial point $\boldsymbol{x}_1$, local step size $\eta_l$, global stepsize $\eta$, server concurrency $M_c$, server updates after receive $M$ updates from clients

```
Initialize sampled set with |M_1| = M_c clients, send server initial model x_1 to active clients
repeat
    Initialize S_t = ∅, clients in M_t perform local SGD updates based on Algorithm 3
    if receive client update then
        Server receive client update Δ^{i_t}_{t−τ^i_t} from client i_t: Δ_t ← Δ_t + Δ^{i_t}_{t−τ^i_t}
        m ← m + 1, S_t ← S_t ∪ {i_t}
    end if
    if m = M then
        Update global model x_{t+1} = x_t + ηΔ_t
        Reset m ← 0, t ← t + 1
        Sample client S_{t+1} ⊆ [N]/M_t ∪ S_t, update M_{t+1} ← M_t/S_t ∪ S_{t+1}, and broadcast global model x_{t+1} to clients
        in S_{t+1}
    end if
until Convergence
```

**Algorithm 3** Asynchronous FL - client

**Input:** Server model (with delay) $\boldsymbol{x}_{t-\tau_t^i}$, learning rate $\eta_l$, number of local SGD steps $K$

```
x^i_{t−τ^i_t,0} = x_{t−τ^i_t}
for k = 0, ..., K − 1 do
    Compute local stochastic gradient: g^i_{t−τ^i_t,k} = ∇F_i(x^i_{t−τ^i_t,k}; ξ)
    x^i_{t−τ^i_t,k+1} = x^i_{t−τ^i_t,k} − η_l g^i_{t−τ^i_t,k}
end for
Obtain model update difference Δ^i_{t−τ^i_t} = x^i_{t−τ^i_t,K} − x_{t−τ^i_t}
```

This also implies the $L$-gradient Lipschitz condition, i.e., $\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$. Assumption B.1 is a standard assumption in nonconvex optimization problems, which has been also adopted in several works (Kingma & Ba, 2015; Reddi et al., 2018; Li et al., 2019; Yang et al., 2021).

**Assumption B.2** (Bounded Variance)**.** Each stochastic gradient on the $i$-th worker 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 B.2 is widely used in federated optimization problems (Li et al., 2019; Reddi et al., 2021; Yang et al., 2021). The bounded local variance represents the randomness of stochastic gradients, and the bounded global variance represents data heterogeneity between clients. Note that $\sigma_g = 0$ corresponds to the *i.i.d* setting, in which datasets from each client have the same distribution.

**Assumption B.3** (Bounded Gradient Delay)**.** Let $\tau_t^i$ represent the delay for global round $t$ and client $i$ which is applied in Algorithm 2 and 3. $\tau_t^i$ implies 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 is bounded, i.e., $\tau_{\max} = \max_{t\in[T], i\in\mathcal{S}_t}\{\tau_t^i\} < \infty$.

Assumption B.3 is a common assumption in convergence analysis for asynchronous federated learning method (Koloskova et al., 2022; Yang et al., 2020). In the following, we will show the convergence results general Asynchronous FL methods.

**Assumption B.4** (Bounded State Delay)**.** Let $\zeta_t^j$ represent the delay of the state variable for global round $t$ and client $j \notin \mathcal{S}_t$ in Algorithm 1. $\zeta_t^j$ is state in the context of client $j$ which does not update the model difference in round $t$ and then maintains the state variable $\boldsymbol{h}_t^j$ as the last step, and $\zeta_t^j$ implies the difference between the current global round $t$ and the global round at which this client $j$ started to compute the last gradient. We assume that the maximum gradient delay is also bounded, i.e., $\zeta_{\max} = \max_{t\in[T], j\notin\mathcal{S}_t}\{\zeta_t^j\} < \infty$.

Assumption B.4 is also commonly used in convergence analysis for asynchronous federated learning method (Koloskova et al., 2022; Yang et al., 2022). In the following, we will show the convergence results for Asynchronous FL and our proposed CA$^2$FL.

**Theorem B.5** (Convergence analysis for Algorithm 2)**.** *Under Assumptions B.1-B.4, if the local learning rate $\eta_l$ and global learning rate $\eta$ satisfy the following condition: $\eta_l \leq \big(\sqrt{\frac{36\eta^2K^2L^2(N-M)^2}{M^2(N-1)^2} - 480K^2L^2\tau_{\max}} - \frac{6\eta KL(N-M)}{M(N-1)}\big)(240K^2L^2\tau_{\max})^{-1}$, and $\eta_l \leq \frac{\eta M(N-1)}{2KN(M-1)}$, then the global rounds of Algorithm 2 satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \frac{1}{\eta\eta_l KT}[f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{t+1})]] + L^2 5K\eta_l^2(\sigma^2 + 6K\sigma_g^2)$$
$$+ \tau_{\max}^2\eta^2\bigg\{\frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{M(N-1)}[15K^3L^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 90K^4L^2\eta_l^2 + 3K^2\sigma_g^2]\bigg\}$$
$$+ \frac{\eta L}{2}\bigg\{\frac{\eta_l}{M}\sigma_l^2 + \frac{\eta_l(N-M)}{M(N-1)}[15K^2TL^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 90K^3L^2\eta_l^2 + 3K\sigma_g^2]\bigg\} \tag{B.1}$$

**Corollary B.6.** *If we choose the local learning rate $\eta = \Theta(\sqrt{KM})$ and $\eta_l = \Theta(1/\sqrt{T}K)$ then the global rounds of Algorithm 2 satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] = \mathcal{O}\bigg(\frac{[(f_0 - f_*) + \sigma^2]}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2 + K\sigma_g^2}{TK}\bigg)$$
$$+ \mathcal{O}\bigg(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2\bigg) + \mathcal{O}\bigg(\frac{K\tau_{\max}^2\sigma_g^2 + \tau_{\max}^2\sigma^2}{T}\bigg), \tag{B.2}$$

*where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$.*

**Theorem B.7** (Convergence analysis for Algorithm 1)**.** *Under Assumptions B.1-B.4, if the local learning rate $\eta_l$ and global learning rate $\eta$ satisfy the following condition: $\eta\eta_l \leq \big(\sqrt{1 + 24\tau_{\max}^2 + \frac{48(N-M)^2\zeta_{\max}^2}{N^2}} - 1\big)\big(12K^2L^2\tau_{\max}^2 + \frac{24K^2L^2(N-M)^2\zeta_{\max}^2}{N^2}\big)^{-1}$ and $\eta_l \leq \big[\big(3\tau_{\max} + \frac{6(N-M)^2\zeta_{\max}}{N^2}\big)2\sqrt{30}KL\big]^{-1}$, then the global rounds of CA$^2$FL satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \frac{1}{\eta\eta_l KT}[f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{T+1})]] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2 + 6K\sigma_g^2)$$
$$+ \frac{3\eta\eta_l L}{2M}\sigma^2 + \bigg(3L^2\tau_{\max}^2 + \frac{6L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{3\eta^2\eta_l^2K\sigma^2}{M}. \tag{B.3}$$

**Corollary B.8.** *If we choose the local learning rate $\eta = \Theta(\sqrt{KM})$ and $\eta_l = \Theta(1/\sqrt{T}K)$ then the global rounds of Algorithm 1 satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] = \mathcal{O}\bigg(\frac{f_0 - f_*}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2 + K\sigma_g^2}{TK}\bigg)$$
$$+ \mathcal{O}\bigg(\frac{\tau_{\max}^2\sigma^2}{T}\bigg) + \mathcal{O}\bigg(\frac{\zeta_{\max}^2(N-M)^2\sigma^2}{TN^2}\bigg), \tag{B.4}$$

*where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$.*

## C. Theoretical Analysis for Memory Friendly Cached-Aided Asynchronous FL

First, we state two additional assumptions for analyzing the quantization method.

**Assumption C.1.** Assume that the random quantization operation $\mathcal{Q}(\boldsymbol{x})$ is unbiased and has bounded variance, i.e.,

$$\mathbb{E}[\mathcal{Q}(\boldsymbol{x})] = \boldsymbol{x}, \quad \mathbb{E}[\|\mathcal{Q}(\boldsymbol{x}) - \boldsymbol{x}\|^2] \leq q\|\boldsymbol{x}\|^2. \tag{C.1}$$

This assumption is a common assumption for quantization methods, which has been adopted in many communication-compression strategies (Reisizadeh et al., 2020; Haddadpour et al., 2021; Alistarh et al., 2017).

**Assumption C.2** (Compression Dissimilarity)**.** For the quantization operator satisfies there exists a constant $\gamma$ such that, for each iteration $t \geq 0$, we have

$$\bigg\| \mathcal{Q}\bigg(\frac{1}{N}\sum_{i=1}^{N}\Delta_t^i\bigg) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{Q}(\Delta_t^i)\bigg\| \leq \gamma \bigg\|\frac{1}{N}\sum_{i=1}^{N}\Delta_t^i\bigg\|. \tag{C.2}$$

Assumption C.2 bounds the difference between the average of compression and compression of average. Similar assumptions have been adopted in Haddadpour et al. (2021).

**Theorem C.3** (Convergence analysis for MF-CA$^2$FL)**.** *Under Assumptions B.1-B.4, Assumptions C.1 and Assumptions C.2, if the local learning rate $\eta_l$ and global learning rate $\eta$ satisfy the following condition: $\eta\eta_l \leq \big(\sqrt{1+\frac{48\tau_{\max}^2}{\gamma^2+q^2}+\frac{96(N-M)^2\zeta_{\max}^2}{N^2(\gamma^2+q^2)}}-1\big)\big(12K^2L^2\tau_{\max}^2+\frac{24K^2L^2(N-M)^2\zeta_{\max}^2}{N^2}\big)^{-1}$ and $\eta_l \leq \big[\big(3\tau_{\max}+\frac{6(N-M)^2\zeta_{\max}}{N^2}\big)2\sqrt{30}KL\big]^{-1}$, then the global rounds of MF-CA$^2$FL satisfy*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &\leq \frac{1}{\eta\eta_l KT}[f(\boldsymbol{x}_1)-\mathbb{E}[f(\boldsymbol{x}_{T+1})]] + \bigg(3+\frac{6(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)\\ &+\frac{3\eta\eta_l L(\gamma^2+q^2)}{M}\sigma^2 + \bigg(3L^2\tau_{\max}^2+\frac{6L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{6\eta^2\eta_l^2K(\gamma^2+q^2)\sigma^2}{M}.\end{aligned} \tag{C.3}$$

**Corollary C.4.** *If we choose the local learning rate $\eta = \Theta(\sqrt{KM})$ and $\eta_l = \Theta(1/\sqrt{T}K)$ then the global rounds of MF-CA$^2$FL satisfy*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &= \mathcal{O}\bigg(\frac{f_0-f_*}{\sqrt{TKM}}\bigg)+\mathcal{O}\bigg(\frac{(\gamma^2+q^2)\sigma^2}{\sqrt{TKM}}\bigg)+\mathcal{O}\bigg(\frac{\sigma^2+K\sigma_g^2}{TK}\bigg)\\ &+\mathcal{O}\bigg(\frac{\tau_{\max}^2(\gamma^2+q^2)\sigma^2}{T}\bigg)+\mathcal{O}\bigg(\frac{\zeta_{\max}^2(N-M)^2(\gamma^2+q^2)\sigma^2}{TN^2}\bigg),\end{aligned} \tag{C.4}$$

*where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$.*

## D. Additional Experiments

In this section, we present additional empirical results for our proposed methods in training CNN network as in (Wang & Ji, 2022) on CIFAR-10, and ResNet-18 network (He et al., 2016) on CIFAR-10/100 (Krizhevsky et al., 2009) datasets, and abundant ablations and discussions about our proposed methods. All experiments in this paper are conducted on 4 NVIDIA RTX A6000 GPUs.

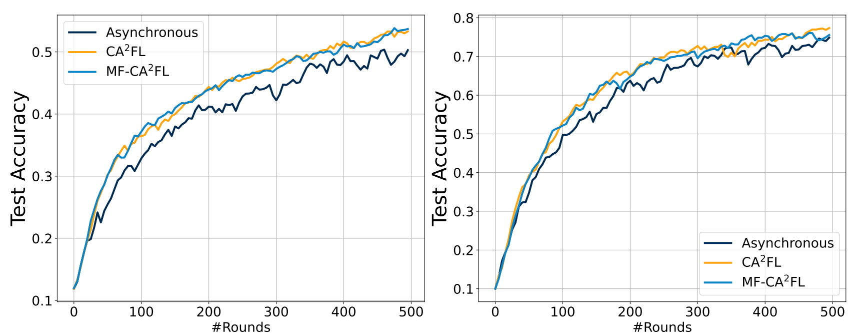
**Implementation overview.** The number of local training iterations $K$ on each client is set to two local epochs (the amount of iteration depends on the amount of data for each client, and and the batch size is set to 50 for all experiments by default. For local update, we use the SGD optimizer with a learning rate gridding from $\{0.01, 0.1, 1\}$ and a global learning rate gridding from $\{0.1, 1\}$. For a fair comparison, the local SGD updates apply no momentum and no gradient clipping steps for all methods. We set a total of 100 clients in the network and the concurrency $M_c = 20$ if there is no further instructions, and we set the update accumulation amount $M = 10$ by default. We simulate the delay distribution from several half-normal distributions similar to FedBuff (Nguyen et al., 2022) controlled by the scaling $\sigma_h$, where larger $\sigma_h$ means in expectation larger wall-clock delay, we default set the half-normal distribution to $\sigma_h \sim$ halfnorm($s$) , where $s \sim$ Unif (0,5).

### D.1. Additional Experimental Results

**Results on CIFAR-10.** Table 2 shows the overall test accuracy of experiments on CIFAR-10 on training different models with two data heterogeneity levels. It demonstrates that our proposed CA$^2$FL achieve better test accuracy than general asynchronous federated learning baselines. Particularly, when the data is highly heterogeneously distributed across clients, indicated by smaller $\alpha$ values in Dirichlet sampling strategies, our CA$^2$FL method significantly outperforms the general asynchronous baseline. Particularly, when $\alpha = 0.1$, CA$^2$FL can significantly outperform Asynchronous FL with more than a 6% increase. Moreover, in the memory-friendly version MF-CA$^2$FL, which reduces the memory overhead by keeping the quantized cached update, the superior performance of the cached variable is still observed and leading to better test accuracy than the general asynchronous baseline. Furthermore, Figure 2 provides the test accuracy curves of training CNN and ResNet-18 networks on CIFAR-10 with $\alpha = 0.3$, offering a visual illustration of the effectiveness of our proposed method.

*Table 2.* The test accuracy of different models on the CIFAR-10 dataset with different models and data heterogeneity degrees. We report the mean accuracy and the standard derivation over 3 runs with different random seeds.

| Method | Dir(0.3) | | Dir(0.1) | |
|---|---|---|---|---|
| | CNN Acc. & std | ResNet-18 Acc. & std | CNN Acc. & std | ResNet-18 Acc. & std |
| Asynchronous FL | $50.15 \pm 1.50$ | $75.60 \pm 1.13$ | $43.71 \pm 3.13$ | $57.31 \pm 4.23$ |
| CA$^2$FL | **53.30** $\pm 0.49$ | **76.36** $\pm 0.57$ | **50.13** $\pm 1.10$ | **68.37** $\pm 1.97$ |
| MF-CA$^2$FL (8 bits) | $52.73 \pm 0.59$ | $74.77 \pm 0.45$ | $49.72 \pm 0.99$ | $67.75 \pm 3.26$ |
| MF-CA$^2$FL (4 bits) | $52.72 \pm 0.45$ | $74.30 \pm 0.73$ | $49.92 \pm 0.62$ | $68.79 \pm 2.82$ |




(a) CNN (b) ResNet-18

*Figure 2.* The test accuracy for our proposed CA$^2$FL and MF-CA$^2$FL (4 bits) with asynchronous federated learning baseline in training CIFAR10 data on CNN and ResNet-18 model.

**Results on CIFAR-100.** Table 3 presents the overall test accuracy of experiments on CIFAR-100 with two data heterogeneity levels. It demonstrates that our proposed CA$^2$FL achieve higher test accuracy compared to the general asynchronous federated learning baseline. Specifically, when the data is highly heterogeneously distributed, e.g., $\alpha = 0.01$, our CA$^2$FL method significantly outperforms the general asynchronous baseline with approximately 4.5% improvement compared to Asynchronous FL. The memory-friendly version MF-CA$^2$FL also shows its advantage over the general asynchronous federated learning baseline.

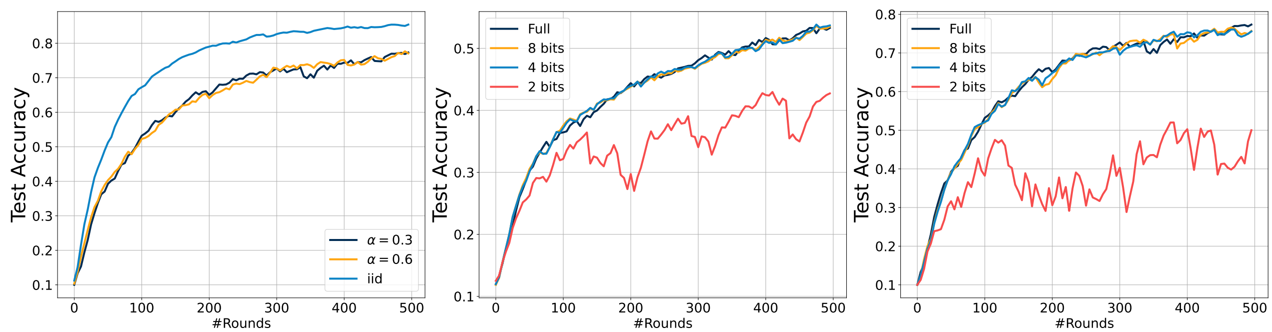
*Table 3.* The test accuracy of different models on the CIFAR-10 dataset with different data heterogeneity degrees. We report the mean accuracy and the standard derivation over 3 runs with different random seeds.

| Method | Dir(0.1) Acc. & std | Dir(0.01) Acc. & std |
|---|---|---|
| Asynchronous FL | $43.64 \pm 1.42$ | $22.15 \pm 1.54$ |
| CA$^2$FL | **44.40** $\pm 1.27$ | **26.67** $\pm 2.20$ |
| MF-CA$^2$FL (8 bits) | $43.84 \pm 0.47$ | $25.89 \pm 0.82$ |
| MF-CA$^2$FL (4 bits) | $43.85 \pm 0.44$ | $25.09 \pm 1.75$ |

### D.2. Ablation Studies and Additional Results

We conduct ablation studies to investigate the effect of maximum asynchronous delay $\tau_{\max}$, the effect of data heterogeneity, the delay sampling strategies, and how different quantization levels would affect the overall convergence and generalization performances of MF-CA$^2$FL. Figure 3 shows curves of test accuracy for different ablations studies.

We show the different levels of data heterogeneity in the plot (a), Figure 3. We also investigate the level of quantization for the proposed MF-CA$^2$FL. From plots (b) and (c), we observe that MF-CA$^2$FL does not suffer from significant performance reduction when quantizing the cached update from the full tensor to 8 bits or 4 bits quantized tensor. This suggests that in CA$^2$FL, the server can save up to $8\times$ storage overhead while still applying the cached update for global model calibration.




(a) Data heterogeneity (b) Quantization with CNN (c) Quantization with ResNet-18

*Figure 3.* Ablation study with several context: (a) the effect of data heterogeneity in CA$^2$FL; different quantization levels on (b) CNN and (c) ResNet-18.

**Concurrency $M_c$, update accumulation $M$ and delay.** Note that both Algorithm 2 and 1 do not explicitly include the delay factor $\tau_{\max}$, and we emphasize that $\tau_{\max}$ is only needed for theoretical analysis. In practice, the delay is controlled by the concurrency $M_c$ and the amount of the model update accumulation $M$ (i.e., the server accumulates model update difference from $M$ different client to update the global model in a round). When the concurrency $M_c$ is large, indicating a large number of clients actively receive the global model from the server and compute the gradient simultaneously, thus if the accumulation number $M$ is small, the delay for clients would be large. Specifically, for plots (a) and (b) in Figure 4, the maximum asynchronous delay $\tau_{\max} = 3, 4, 5$ correspond to the pairs of $M_c = 15, M = 10$, $M_c = 20, M = 10$ and $M_c = 25, M = 10$, i.e., this $\tau_{\max}$ ablation is the same as the ablation study of network concurrency $M_c$.

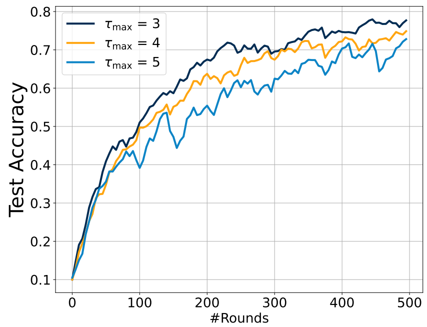
Moreover, another ablation study regarding $\tau_{\max}$ is the ablation for the amount of the model update accumulation $M$. Figure 4 shows the test accuracy for three pairs: $M_c = 20, M = 15$, $M_c = 20, M = 10$ and $M_c = 20, M = 5$, with the corresponding $\tau_{\max} = 2, 4, 9$. This shows that the maximum delay $\tau_{\max}$ could also be varying from different model update accumulation $M$ given the same concurrency $M_c$. We further track the average delay

$$\tau_{avg} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}\tau_t^i \tag{D.1}$$

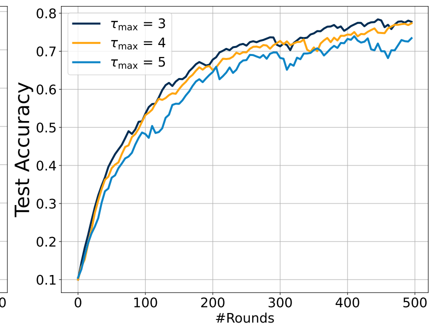
of asynchronous algorithms, which might better describe the delay in both Asynchronous FL and our proposed CA$^2$FL. We summarize the maximum asynchronous and average delay for the experiments shown in Figure 4 and Table 4.

**Simulated delay distributions.** We sample the wall-clock delay distributions from several half-normal distributions. We have investigated a different delay distribution simulation strategy in the plot (a), Figure 5. It shows that if all clients' wall-clock delays are sampled from the same halfnorm(5) distribution, i.e., $\sigma_h \sim$ halfnorm(5), the overall performance would be a little worse than each client's wall-clock time delay is sampled from a random half-normal distribution, i.e., $\sigma_h \sim$ halfnorm($s$) and $s \sim$ Unif(0, 5). It is worth mentioning that the fixed half-normal distribution leads to $\tau_{\max} = 2$ and $\tau_{\max} = 0.9946$, and random half-normal distribution leads to $\tau_{\max} = 4$ and $\tau_{\max} = 0.9184$. This further demonstrates that the average delay is also important for the overall performance of the asynchronous federated learning methods.

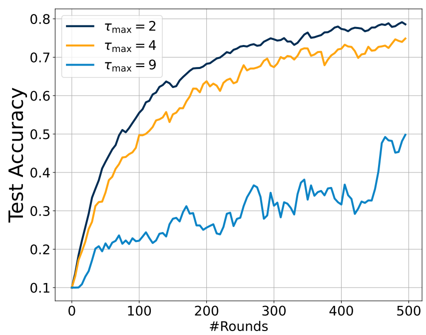
We compare how the parameter of the half-normal distribution would affect the overperformance of our proposed method. We compare experiments with $\sigma_h \sim$ halfnorm($s$) and $s \sim$ Unif(0, 5), $\sigma_h \sim$ halfnorm($s$) and $s \sim$ Unif(0, 1) and $\sigma_h \sim$



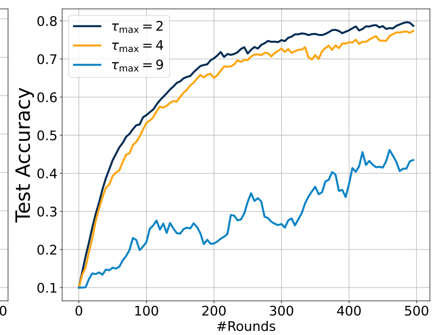

(a) $\tau_{\max}$ in Async FL




(b) $\tau_{\max}$ in CA$^2$FL




(c) Asynchronous FL




(d) CA$^2$FL

*Figure 4.* Ablation study with several contexts: (a) the effect of maximum delay in the Asynchronous FL baseline; (b) the effect of maximum delay in the proposed CA$^2$FL; different maximum asynchronous delay by adopting different accumulation amount $M$ in (c) the Asynchronous FL baseline and (d) the proposed CA$^2$FL when training ResNet-18 network on CIFAR-10 dataset.

halfnorm($s$) and $s \sim$ Unif(0, 0.5), which implies that the parameter for wall-clock delays are randomly sampled from different uniform distributions. Figure 5 plots (b) and (c) show that there is only a slight difference between several uniform distributions. We track the maximum and average asynchronous delay and we summarize these numerical factors in Table 5. It shows that although the maximum delay differs from uniform distributions, the average delay is very similar, thus the overall performance is similar when choosing different uniform distributions for simulating the wall-clock delay.

**Comparison with FedAsync (Xie et al., 2019).** FedAsync (Xie et al., 2019) is one of the first works studying asynchronous federated learning methods in which clients jointly train a model with their local private data at their own pace. In FedAsync, each client conducts $K$ steps of local SGD training to solve a regularized optimization problem, i.e,

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{E}_{\xi\sim\mathcal{D}_i}[F_i(\boldsymbol{x};\xi_i)] + \frac{\rho}{2}\|\boldsymbol{x}-\boldsymbol{x}_t\|^2. \tag{D.2}$$

Once the server receives a local model from an arbitrary client, it would immediately update the global model by adopting a momentum average strategy:

$$\boldsymbol{x}_{t+1} = (1-\alpha_t)\boldsymbol{x}_t + \alpha_t \boldsymbol{x}^i_{t-\tau,K}. \tag{D.3}$$

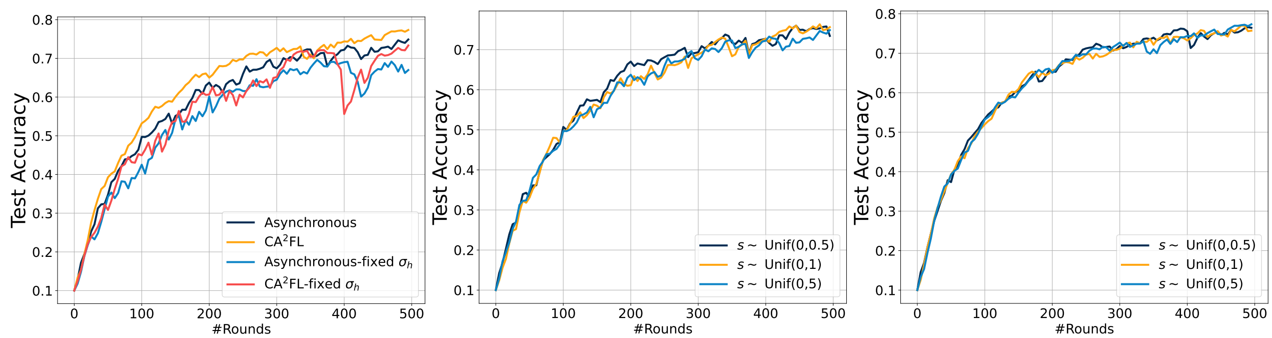
The momentum factor $\alpha_t$ can be updated in various ways, but we specifically compare our method with two variants: 1) constant update: $\alpha_t = \alpha \cdot s(t-\tau)$, where $s(t-\tau) = 1$, and 2) polynomial update: $\alpha_t = \alpha \cdot s(t-\tau)$, where $s(t-\tau) = (t-\tau+1)^{-\beta}$ with $\beta > 0$. Figure 6 illustrates that our proposed CA$^2$FL achieves significantly better test accuracy results compared to FedAsync with these two different momentum averaging strategies. This further demonstrates the superior performance of our cached-aided asynchronous federated learning method.

### D.3. Hyper-parameters Details

We conduct detailed hyper-parameter searches to find the best hyper-parameter for each baseline. We grid over the local learning rater $\eta_l \in \{0.001, 0.01, 0.1, 1.0\}$, and the global learning rate $\eta \in \{0.001, 0.01, 0.1, 1.0, 2.0\}$ for each methods.

*Table 4.* The maximum and average with different concurrency and clients' update accumulation when training ResNet-18 network on CIFAR-10 dataset.

| Settings | $M_c = 15, M = 10$ | $M_c = 20, M = 10$ | $M_c = 25, M = 10$ |
|---|---|---|---|
| $\tau_{\max}$ | 3 | 4 | 5 |
| $\tau_{avg}$ | 0.4888 | 0.9184 | 1.312 |
| Settings | $M_c = 20, M = 15$ | $M_c = 20, M = 10$ | $M_c = 20, M = 5$ |
| $\tau_{\max}$ | 2 | 4 | 9 |
| $\tau_{avg}$ | 0.3264 | 0.9181 | 2.6512 |



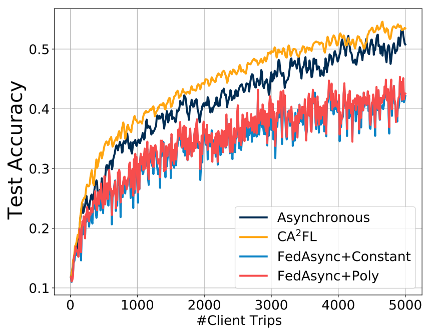

(a) Delay simulations (b) Asynchronous FL (c) CA$^2$FL

*Figure 5.* Ablation study for different distributions for simulating the wall-clock delay.

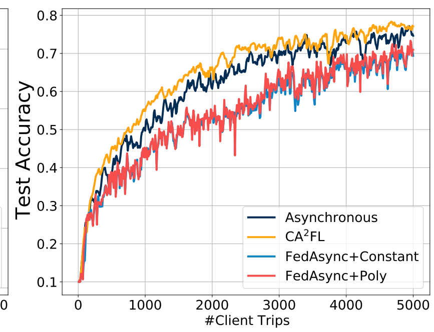
Table 6 summarizes the hyper-parameter details in our experiments. Experiments are set up with 100 total clients, the concurrency is $M_c = 20$ by default, and we let the server update the global model once it receives $M = 10$ updates from clients. For each method, we conduct 2 local epochs (the explicit local iterations $K$ may differ from clients) of local training with a batch size of 50 by default. We set the weight decay as $10^{-4}$ for the local SGD optimizer. For FedAsync (Xie et al., 2019), we additionally grid over the weight of the regularization term $\rho \in \{0.01, 0.1, 1.0\}$, the momentum factor $\alpha_t \in \{0.1, 0.3, 0.5, 0.9\}$, and $\beta \in \{0.3, 0.5\}$ in the polynomial update. Table 7 presents the hyper-parameter details of FedAsync (Xie et al., 2019).

*Table 5.* The maximum and average delay corresponding to ablations in Figure 5, plot (b) and plot (c).

| | $s \sim$ Unif(0, 0.5) | $s \sim$ Unif(0, 1) | $s \sim$ Unif(0, 5) |
|---|---|---|---|
| $\tau_{\max}$ | 6 | 6 | 4 |
| $\tau_{avg}$ | 0.8992 | 0.9092 | 0.9184 |




(a) CNN




(b) Resnet-18

*Figure 6.* Comparison with FedAsync when training CNN and ResNet-18 model on CIFAR-10 dataset.

*Table 6.* Hyper-parameters details.

| CIFAR-10 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Models & Dir($\alpha$) | Asynchronous FL $\eta_l$ | $\eta$ | CA$^2$FL $\eta_l$ | $\eta$ | MF-CA$^2$FL (8 bits) $\eta_l$ | $\eta$ | MF-CA$^2$FL (4 bits) $\eta_l$ | $\eta$ |
| CNN & Dir(0.3) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |
| ResNet-18 & Dir(0.3) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |
| CNN & Dir(0.1) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |
| ResNet-18 & Dir(0.1) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |
| CIFAR-100 | | | | | | | | |
| Models & Dir($\alpha$) | Asynchronous FL $\eta_l$ | $\eta$ | CA$^2$FL $\eta_l$ | $\eta$ | MF-CA$^2$FL (8 bits) $\eta_l$ | $\eta$ | MF-CA$^2$FL (4 bits) $\eta_l$ | $\eta$ |
| ResNet-18 & Dir(0.1) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |
| ResNet-18 & Dir(0.01) | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 | 0.01 | 1.0 |

*Table 7.* Additional hyper-parameters of FedAsync (Xie et al., 2019).

| CIFAR-10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models & Dir($\alpha$) | Constant update $\eta_l$ | $\rho$ | $\alpha_t$ | Polynomial update $\eta_l$ | $\rho$ | $\alpha_t$ | $\beta$ |
| CNN & Dir(0.3) | 0.01 | 1.0 | 0.3 | 0.01 | 1.0 | 0.5 | 0.3 |
| ResNet-18 & Dir(0.3) | 0.01 | 0.1 | 0.1 | 0.01 | 0.1 | 0.3 | 0.3 |

## E. Convergence Analysis for Asynchronous FL

*Proof of Theorem B.5.* Since $f$ is $L$-smooth, taking conditional expectation at time $t$, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1})] - f(\boldsymbol{x}_t) \\
&\leq \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] \\
&= \underbrace{\mathbb{E}[\langle \nabla f(\boldsymbol{x}_t)), \eta\Delta_t \rangle]}_{I} + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}[\|\Delta_t\|^2]}_{II}. \qquad \text{(E.1)}
\end{aligned}
$$

**Bounding $I_1$**

$$
\begin{aligned}
I_1 &= \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \eta\Delta_t \rangle] \\
&= \frac{\eta}{M}\sum_{i\in\mathcal{M}_t} \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \Delta^i_{t-\tau^i_t} \rangle] \\
&= -\frac{\eta\eta_l}{M}\sum_{i\in\mathcal{M}_t} \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\tau^i_t,k} \bigg\rangle\bigg] \\
&= -\frac{\eta\eta_l}{M}\sum_{i\in\mathcal{M}_t}\sum_{k=0}^{K-1} \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{g}^i_{t-\tau^i_t,k} \rangle] \\
&= -\frac{\eta\eta_l}{M}\sum_{i\in\mathcal{M}_t}\sum_{k=0}^{K-1} \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \rangle] \\
&= -\eta\eta_l \sum_{k=0}^{K-1} \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\rangle\bigg], \qquad \text{(E.2)}
\end{aligned}
$$

where the second and third equation holds by the update rule. The fifth one holds by the unbiasedness of stochastic gradient. 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

$$
\begin{aligned}
&- \eta\eta_l \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\rangle\bigg] \\
&= -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2K}\mathbb{E}\bigg[\bigg\| \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\|^2\bigg] \\
&+ \frac{\eta\eta_l}{2}\mathbb{E}\bigg[\bigg\| \sqrt{K}\nabla f(\boldsymbol{x}_t) - \frac{1}{N\sqrt{K}}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\|^2\bigg] \\
&= -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2K}\mathbb{E}\bigg[\bigg\| \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\|^2\bigg] \\
&+ \frac{\eta\eta_l}{2}\sum_{k=0}^{K-1}\mathbb{E}\bigg[\bigg\| \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\boldsymbol{x}_t) - \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \bigg\|^2\bigg], \qquad \text{(E.3)}
\end{aligned}
$$

for the last term, we have

$$\begin{aligned}
&\frac{\eta\eta_l}{2}\sum_{k=0}^{K-1}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_t)-\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg]\\
&\leq \frac{\eta\eta_l}{2}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2]\\
&\leq \frac{\eta\eta_l}{N}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\Big[\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\tau^i_t})\|^2]+\mathbb{E}[\|\nabla F_i(\boldsymbol{x}_{t-\tau^i_t})-\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2]\Big]\\
&\leq \frac{\eta\eta_l}{N}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\Big[L^2\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}_{t-\tau^i_t}\|^2]+L^2\mathbb{E}[\|\boldsymbol{x}_{t-\tau^i_t}-\boldsymbol{x}^i_{t-\tau^i_t,k}\|^2]\Big]. \qquad (E.4)
\end{aligned}$$

For the first term, we have

$$\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}_{t-\tau^i_t}\|^2]=\mathbb{E}\bigg[\bigg\|\sum_{s=t-\tau^i_t}^{t-1}(\boldsymbol{x}_{s+1}-\boldsymbol{x}_s)\bigg\|^2\bigg]\leq\tau_{\max}\sum_{s=t-\tau^i_t}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2] \qquad (E.5)$$

For the second term, we have

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

Thus we have

$$\begin{aligned}
&\frac{\eta\eta_l}{2}\sum_{k=0}^{K-1}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_t)-\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg]\\
&\leq \eta\eta_l KL^2[5K\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]]+\eta\eta_l KL^2\tau_{\max}\sum_{s=t-\tau^i_t}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2]. \qquad (E.7)
\end{aligned}$$

Thus for $I_1$, we have

$$\begin{aligned}
I_1 &\leq -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]-\frac{\eta\eta_l}{2K}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg]\\
&\quad+\eta\eta_l KL^2[5K\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]]+\eta\eta_l KL^2\tau_{\max}\sum_{s=t-\tau^i_t}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2]. \qquad (E.8)
\end{aligned}$$

**Bounding $I_2$**

$$\begin{aligned}
I_2 &= \frac{\eta^2 L}{2}\mathbb{E}[\|\Delta_t\|^2]=\frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{M}_t}\Delta^i_{t-\tau^i_t}\bigg\|^2\bigg]\\
&\leq \frac{\eta^2 L}{2}\bigg\{\frac{K\eta_l^2}{M}\sigma_l^2+\frac{\eta_l^2(N-M)}{NM(N-1)}[15NK^3L^3\eta_l^2(\sigma_l^2+6K\sigma_g^2)+90NK^4L^2\eta_l^2+3NK^2\|\nabla f(\boldsymbol{x}_t)\|^2\\
&\quad+3NK^2\sigma_g^2]+\frac{\eta_l^2(M-1)}{NM(N-1)}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\bigg\|^2\bigg]\bigg\}. \qquad (E.9)
\end{aligned}$$

For simplicity, in the following, we define $\mathbf{V}_t = \sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i)$.

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

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{t+1})] - f(\mathbf{z}_t) = I_1 + I_2 + I_3 \\
&\leq -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \frac{\eta\eta_l}{2K}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i)\bigg\|^2\bigg] \\
&+ \eta\eta_l K L^2[5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]] + \eta\eta_l K L^2\tau_{\max}\sum_{s=t-\tau_t^i}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2] \\
&+ \frac{\eta^2 L}{2}\bigg\{\frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}\bigg[15NK^3L^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 90NK^4L^2\eta_l^2 + 3NK^2\|\nabla f(\boldsymbol{x}_t)\|^2 \\
&+ 3NK^2\sigma_g^2\bigg] + \frac{\eta_l^2(M-1)}{NM(N-1)}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t-\tau_t^i,k}^i)\bigg\|^2\bigg]\bigg\} \\
&\leq -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \eta\eta_l K L^2[5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]] \\
&+ \eta\eta_l K L^2\tau_{\max}\sum_{s=t-\tau_t^i}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2] + \frac{\eta^2 L}{2}\bigg\{\frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}\bigg[15NK^3L^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) \\
&+ 90NK^4L^2\eta_l^2 + 3NK^2\|\nabla f(\boldsymbol{x}_t)\|^2 + 3NK^2\sigma_g^2\bigg] + \bigg(\frac{\eta_l^2(M-1)}{NM(N-1)} - \frac{\eta\eta_l}{2KN^2}\bigg)\mathbb{E}[\|\mathbf{V}_t\|^2]\bigg\}.
\end{aligned} \tag{E.10}
$$

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

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \eta\eta_l K L^2[5K\eta_l^2 T(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]] \\
&+ \eta\eta_l K L^2\tau_{\max}\sum_{t=1}^{T}\sum_{s=t-\tau_t^i}^{t-1}\mathbb{E}[\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2] + \frac{\eta^2 L}{2}\bigg\{\frac{KT\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}\bigg[15NK^3TL^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) \\
&+ 90NK^4TL^2\eta_l^2 + 3NK^2\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + 3NK^2T\sigma_g^2\bigg] + \bigg(\frac{\eta_l^2(M-1)}{NM(N-1)} - \frac{\eta\eta_l}{2KM^2}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2]\bigg\} \\
&\leq -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \eta\eta_l KTL^2[5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) + 30K^2\eta_l^2\tau_{\max}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]] \\
&+ \eta\eta_l K L^2\tau_{\max}^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] + \frac{\eta^2 L}{2}\bigg\{\frac{KT\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}\bigg[15NK^3TL^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) \\
&+ 90NK^4TL^2\eta_l^2 + 3NK^2\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + 3NK^2T\sigma_g^2\bigg] + \bigg(\frac{\eta_l^2(M-1)}{NM(N-1)} - \frac{\eta\eta_l}{2KN^2}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2]\bigg\}.
\end{aligned} \tag{E.11}
$$

thus by the constraint as follows,

$$\begin{aligned}
& \frac{\eta_l^2(M-1)}{NM(N-1)} - \frac{\eta\eta_l}{2KN^2} \leq 0 \\
\Rightarrow \quad & \eta_l \leq \frac{\eta M(N-1)}{2KN(M-1)}, \\
& \frac{\eta^2 L}{2} \frac{\eta_l^2(N-M)}{NM(N-1)} 3NK^2 + 30\eta\eta_l K L^2 K^2 \eta_l^2 \tau_{\max} \leq \frac{\eta\eta_l K}{4} \\
\Rightarrow \quad & \frac{6\eta\eta_l L(N-M)}{M(N-1)} K + 120K^2\eta_l^2 L^2 \tau_{\max} \leq 1 \\
\Rightarrow \quad & \eta_l \leq \left(\sqrt{\frac{36\eta^2K^2L^2(N-M)^2}{M^2(N-1)^2} - 480K^2L^2\tau_{\max}} - \frac{6\eta K L(N-M)}{M(N-1)}\right)(240K^2L^2\tau_{\max})^{-1},
\end{aligned} \tag{E.12}$$

thus we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq & \frac{1}{\eta\eta_l KT}[f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{t+1})]] + L^2 5K\eta_l^2(\sigma^2 + 6K\sigma_g^2) \\
& + \tau_{\max}^2\eta^2\left\{\frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{M(N-1)}[15K^3L^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 90K^4L^2\eta_l^2 + 3K^2\sigma_g^2]\right\} \\
& + \frac{\eta L}{2}\left\{\frac{\eta_l}{M}\sigma_l^2 + \frac{\eta_l(N-M)}{M(N-1)}[15K^2TL^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 90K^3L^2\eta_l^2 + 3K\sigma_g^2]\right\}
\end{aligned} \tag{E.13}$$

□

By choosing $\eta = \Theta(\sqrt{KM})$ and $\eta_l = \Theta(1/\sqrt{T}K)$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] = & \mathcal{O}\left(\frac{[(f_0 - f_*) + \sigma^2]}{\sqrt{TKM}}\right) + \mathcal{O}\left(\frac{\sigma^2 + K\sigma_g^2}{TK}\right) \\
& + \mathcal{O}\left(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2\right) + \mathcal{O}\left(\frac{K\tau_{\max}^2\sigma_g^2 + \tau_{\max}^2\sigma^2}{T}\right),
\end{aligned} \tag{E.14}$$

where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$.

## F. Convergence Analysis for CA$^2$FL

*Proof of Theorem B.7.* By the update scheme of Algorithm 1, we have

$$
\begin{aligned}
&\boldsymbol{v}_t \leftarrow \boldsymbol{h}_t + \frac{1}{M}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_{t-1}) \Rightarrow \boldsymbol{v}_t = \boldsymbol{h}_{t-1} + \frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_{t-1}).\\
&\boldsymbol{v}_t = \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\boldsymbol{h}^i_{t-1} + \frac{1}{N}\sum_{i\in\mathcal{S}_t}\boldsymbol{h}^i_{t-1} + \frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_{t-1})\\
&\quad = \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\boldsymbol{h}^i_{t-1} + \sum_{i\in\mathcal{S}_t}\left[\left(\frac{1}{N} - \frac{1}{M}\right)\boldsymbol{h}^i_{t-1} + \frac{1}{M}\Delta^i_{t-\tau^i_t}\right]
\end{aligned}
\tag{F.1}
$$

Since $f$ is $L$-smooth, taking conditional expectation at time $t$, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1})] - f(\boldsymbol{x}_t)\\
&\leq \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2]\\
&= \underbrace{\mathbb{E}[\langle \nabla f(\boldsymbol{x}_t)), \eta\boldsymbol{v}_t\rangle]}_{I} + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}[\|\boldsymbol{v}_t\|^2]}_{II}.
\end{aligned}
\tag{F.2}
$$

Since we $\boldsymbol{h}^i_t$ represents the state update for client $i$, and $\boldsymbol{h}^i_t$ keeps unchanged if $i \notin \mathcal{S}_t$. We have the following

$$
\boldsymbol{h}_t = \boldsymbol{h}_{t-1} + \frac{1}{N}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_{t-1}) = \frac{1}{N}\sum_{i\in\mathcal{S}_t}\Delta^i_{t-\tau^i_t} + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\Delta^i_{t-\zeta^i_t},
\tag{F.3}
$$

**Bounding $I$**

$$
\begin{aligned}
I &= \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \eta \boldsymbol{v}_t \rangle] \\
&= \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{\eta}{M} \sum_{i \in \mathcal{S}_t} \Delta^i_{t-\tau^i_t} + \bigg(\frac{\eta}{N} - \frac{\eta}{M}\bigg) \sum_{i \in \mathcal{S}_t} \boldsymbol{h}^i_{t-1} + \frac{\eta}{N} \sum_{i \notin \mathcal{S}_t} \boldsymbol{h}^i_{t-1} \bigg\rangle\bigg] \\
&= -\eta\eta_l \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{1}{M} \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\tau^i_t,k} + \bigg(\frac{1}{N} - \frac{1}{M}\bigg) \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\zeta^i_t,k} + \frac{1}{N} \sum_{i \notin \mathcal{S}_t} \sum_{k=0}^{K-1} \boldsymbol{g}^i_{t-\zeta^i_t,k} \bigg\rangle\bigg] \\
&= -\eta\eta_l \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{1}{M} \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg) \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \\
&\quad + \frac{1}{N} \sum_{i \notin \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \bigg\rangle\bigg] \\
&= -\eta\eta_l K \mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_t), \frac{1}{MK} \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{NK} - \frac{1}{MK}\bigg) \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \\
&\quad + \frac{1}{NK} \sum_{i \notin \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \bigg\rangle\bigg] \\
&= -\frac{\eta\eta_l K}{2} \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&\quad - \frac{\eta\eta_l}{2K} \mathbb{E}\bigg[\bigg\| \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \bigg(\frac{1}{M} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg) \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) + \frac{1}{N} \sum_{i \notin \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \bigg]\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l K}{2} \mathbb{E}\bigg[\bigg\| \nabla f(\boldsymbol{x}_t) - \frac{1}{K}\bigg[ \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \bigg(\frac{1}{M} \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg) \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) \\
&\quad + \frac{1}{N} \sum_{i \notin \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \bigg]\bigg\|^2\bigg],
\end{aligned} \tag{F.4}
$$

where the second and third equation holds by the update rule. The forth one holds by the unbiasedness of stochastic gradient, and the last one holds 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 the last item, we have

$$
\begin{aligned}
&\frac{\eta\eta_l K}{2}\mathbb{E}\bigg[\bigg\|\nabla f(\boldsymbol{x}_t) - \frac{1}{K}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) \\
&+ \sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\frac{1}{N}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg]\bigg\|^2\bigg] \\
=&\frac{\eta\eta_l K}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_t) - \frac{1}{K}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k}) \\
&+ \sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\frac{1}{N}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg]\bigg\|^2\bigg] \\
=&\frac{\eta\eta_l K}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{MK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\tau^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})] + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\frac{1}{K}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i\boldsymbol{x}_{t-\zeta^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})] \\
&+ \frac{1}{NK}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})] + \frac{1}{M}\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\tau^i_t})] \\
&+ \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\zeta^i_t})] + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\zeta^i_t})]\bigg\|^2\bigg] \\
\leq&\eta\eta_l K\mathbb{E}\bigg[\bigg\|\frac{1}{MK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\tau^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})] + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\frac{1}{K}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})] \\
&+ \frac{1}{NK}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta^i_t}) - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})]\bigg\|^2\bigg] + \eta\eta_l K\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\tau^i_t})] \\
&+ \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\zeta^i_t})] + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t) - \nabla F_i(\boldsymbol{x}_{t-\zeta^i_t})]\bigg\|^2\bigg],
\end{aligned}
\tag{F.5}
$$

where we have

$$
\begin{aligned}
&\eta\eta_l K\mathbb{E}\bigg[\bigg\|\frac{1}{MK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\tau_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})]+\bigg(\frac{1}{N}-\frac{1}{M}\bigg)\frac{1}{K}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\zeta_t^i,k})]\\
&\quad+\frac{1}{NK}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\zeta_t^i,k})]\bigg\|^2\bigg]\\
\le&\frac{3\eta\eta_l K}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\bigg\|\frac{1}{K}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\tau_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\tau_t^i,k})]\bigg\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)^2}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\bigg\|\frac{1}{K}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\zeta_t^i,k})]\bigg\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\bigg\|\frac{1}{K}\sum_{k=0}^{K-1}[\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})-\nabla F_i(\boldsymbol{x}^i_{t-\zeta_t^i,k})]\bigg\|^2\bigg]\\
\le&\frac{3\eta\eta_l K}{M}M\cdot[5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]]\\
&\quad+\frac{3\eta\eta_l K(N-M)^2}{N^2M}M\cdot[5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]]\\
&\quad+\frac{3\eta\eta_l K(N-M)}{N^2}(N-M)\cdot[5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]]\\
\le&3\eta\eta_l K\cdot[5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]]\\
&\quad+\frac{6\eta\eta_l K(N-M)^2}{N^2}[5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)+30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]].
\end{aligned}
\tag{F.6}
$$

We also have

$$
\begin{aligned}
&\eta\eta_l K\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\tau_t^i})]+\bigg(\frac{1}{N}-\frac{1}{M}\bigg)\sum_{i\in\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})]\\
&\quad+\frac{1}{N}\sum_{i\notin\mathcal{S}_t}[\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})]\bigg\|^2\bigg]\\
\le&\frac{3\eta\eta_l K}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\tau_t^i})\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\|\nabla F_i(\boldsymbol{x}_t)-\nabla F_i(\boldsymbol{x}_{t-\zeta_t^i})\|^2\bigg]\\
\le&\frac{3\eta\eta_l KL^2\tau_{\max}}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg].
\end{aligned}
\tag{F.7}
$$

**Bounding *II***

$$\begin{aligned}
II &= \frac{\eta^2 L}{2}\mathbb{E}[\|\boldsymbol{v}_t\|^2] = \frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}\Delta^i_{t-\tau^i_t} + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)\sum_{i\in\mathcal{S}_t}\boldsymbol{h}^i_t + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\boldsymbol{h}^i_t\bigg\|^2\bigg] \\
&\le \frac{\eta^2\eta_l^2 L}{2}\mathbb{E}\bigg[\bigg\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}[\boldsymbol{g}^i_{t-\tau^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})] + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)[\boldsymbol{g}^i_{t-\zeta^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})]\bigg) \\
&\quad + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}[\boldsymbol{g}^i_{t-\zeta^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})]\bigg\|^2\bigg] + \frac{\eta^2\eta_l^2 L}{2}\mathbb{E}\bigg[\bigg\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \\
&\quad + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg\|^2\bigg] \\
&\le \frac{\eta^2\eta_l^2 L}{2}\frac{1}{M^2}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\mathbb{E}[\|\boldsymbol{g}^i_{t-\tau^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k})\|^2] + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)^2\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\mathbb{E}[\|\boldsymbol{g}^i_{t-\zeta^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\|^2] \\
&\quad + \frac{1}{N^2}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\mathbb{E}[\|\boldsymbol{g}^i_{t-\zeta^i_t,k} - \nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\|^2] + \frac{\eta^2\eta_l^2 L}{2}\mathbb{E}\bigg[\bigg\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) \\
&\quad + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg\|^2\bigg] \\
&\le \frac{\eta^2\eta_l^2 L}{2}\frac{3K}{M}\sigma^2 + \frac{\eta^2\eta_l^2 L}{2}\mathbb{E}\bigg[\bigg\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N}-\frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) \\
&\quad + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg\|^2\bigg]. \qquad \text{(F.8)}
\end{aligned}$$

**Merging pieces.** For simplicity, we define $\mathbf{V}_t = \sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\big(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \big(\frac{1}{N}-\frac{1}{M}\big)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\big) +$

$\frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})$. Therefore, by merging pieces together, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1})]-f(\boldsymbol{x}_t)=I+II\\
&\leq -\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]-\frac{\eta\eta_l}{2K}\mathbb{E}[\|\mathbf{V}_t\|^2]+\bigg(3\eta\eta_l K+\frac{6\eta\eta_l K(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)\\
&\quad+3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]+\frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]\\
&\quad+\frac{3\eta\eta_l KL^2\tau_{\max}}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{\eta^2L}{2}\bigg\{\frac{3K\eta_l^2}{M}\sigma^2+\eta_l^2\mathbb{E}[\|\mathbf{V}_t\|^2]\bigg\}\\
&=-\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]+\bigg(3+\frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+\eta^2L\frac{3K\eta_l^2}{2M}\sigma^2\\
&\quad+\frac{3\eta\eta_l KL^2\tau_{\max}}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2]+\frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]\\
&\quad-\bigg(\frac{\eta\eta_l}{2K}-\frac{\eta^2\eta_l^2L}{2}\bigg)\mathbb{E}[\|\mathbf{V}_t\|^2].
\end{aligned}
\tag{F.9}
$$

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

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{T+1})] - f(\boldsymbol{x}_1) \\
\leq & -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L\frac{3KT\eta_l^2}{2M}\sigma^2 \\
&+ \frac{3\eta\eta_l K L^2\tau_{\max}}{M}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau_t^i}^{t-1}\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2\bigg] + \frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2 M}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2\bigg] \\
&+ \frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|^2\bigg] \\
&+ 3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] + \frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2] \\
&- \bigg(\frac{\eta\eta_l}{2K} - \frac{\eta^2\eta_l^2 L}{2}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2] \\
\leq & -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L\frac{3KT\eta_l^2}{2M}\sigma^2 \\
&+ \bigg(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] \\
&+ 3\eta\eta_l K\sum_{t=1}^{T}\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] \\
&+ \frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{t=1}^{T}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2] - \bigg(\frac{\eta\eta_l}{2K} - \frac{\eta^2\eta_l^2 L}{2}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2],
\end{aligned} \tag{F.10}$$

while previously we obtained

$$\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] \leq \eta^2\frac{3K\eta_l^2}{M}\sigma^2 + \eta^2\eta_l^2\mathbb{E}[\|\mathbf{V}_t\|^2], \tag{F.11}$$

with the constraint of

$$\begin{aligned}
&\frac{\eta^2\eta_l^2 L}{2} + \eta^2\eta_l^2\bigg(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg) \leq \frac{\eta\eta_l}{2K} \\
\Rightarrow &\eta\eta_l K L + \eta^2\eta_l^2\bigg(6K^2L^2\tau_{\max}^2 + \frac{12K^2L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg) \leq 1 \\
\Rightarrow &\eta\eta_l \leq \bigg(\sqrt{1 + 24\tau_{\max}^2 + \frac{48(N-M)^2\zeta_{\max}^2}{N^2}} - 1\bigg)\bigg(12KL\tau_{\max}^2 + \frac{24KL(N-M)^2\zeta_{\max}^2}{N^2}\bigg)^{-1},
\end{aligned} \tag{F.12}$$

and

$$\begin{aligned}
&\bigg(3\eta\eta_l K\tau_{\max} + \frac{6\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2}\bigg)(30K^2L^2\eta_l^2) \leq \frac{\eta\eta_l K}{4} \\
\Rightarrow &\eta_l \leq \bigg[\bigg(3\tau_{\max} + \frac{6(N-M)^2\zeta_{\max}}{N^2}\bigg)2\sqrt{30}KL\bigg]^{-1},
\end{aligned} \tag{F.13}$$

Thus we have

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{T+1})] - f(\boldsymbol{x}_1) \\
&\leq -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L \frac{3KT\eta_l^2}{2M}\sigma^2 \\
&\quad + \bigg(3\eta\eta_l K L^2 \tau_{\max}^2 + \frac{6\eta\eta_l K L^2 (N-M)^2 \zeta_{\max}^2}{N^2}\bigg)\frac{3\eta^2\eta_l^2 K T \sigma^2}{M} \\
&\quad + \bigg(3\eta\eta_l K \tau_{\max} + \frac{6\eta\eta_l K (N-M)^2 \zeta_{\max}}{N^2}\bigg)(30K^2L^2\eta_l^2)\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\
&\leq -\frac{\eta\eta_l K}{4}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L \frac{3KT\eta_l^2}{2M}\sigma^2 \\
&\quad + \bigg(3\eta\eta_l K L^2 \tau_{\max}^2 + \frac{6\eta\eta_l K L^2 (N-M)^2 \zeta_{\max}^2}{N^2}\bigg)\frac{3\eta^2\eta_l^2 K T \sigma^2}{M}.
\end{aligned} \tag{F.14}$$

Therefore,

$$\begin{aligned}
\frac{\eta\eta_l K}{4}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &\leq f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{T+1})] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) \\
&\quad + \eta^2 L \frac{3KT\eta_l^2}{2M}\sigma^2 + \bigg(3\eta\eta_l K L^2 \tau_{\max}^2 + \frac{6\eta\eta_l K L^2 (N-M)^2 \zeta_{\max}^2}{N^2}\bigg)\frac{3\eta^2\eta_l^2 K T \sigma^2}{M}, \\
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &\leq \frac{1}{\eta\eta_l K T}[f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{T+1})]] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2 + 6K\sigma_g^2) \\
&\quad + \frac{3\eta\eta_l L}{2M}\sigma^2 + \bigg(3L^2\tau_{\max}^2 + \frac{6L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{3\eta^2\eta_l^2 K \sigma^2}{M}.
\end{aligned} \tag{F.15}$$

□

*Proof of Corollary B.8.* Hence by choosing $\eta_l = \frac{1}{\sqrt{TK}}$ and $\eta = \sqrt{KM}$, then the convergence rate satisfies

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &= \mathcal{O}\bigg(\frac{f_0 - f_*}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2 + K\sigma_g^2}{TK}\bigg) \\
&\quad + \mathcal{O}\bigg(\frac{\tau_{\max}^2\sigma^2}{T}\bigg) + \mathcal{O}\bigg(\frac{\zeta_{\max}^2(N-M)^2\sigma^2}{TN^2}\bigg),
\end{aligned} \tag{F.16}$$

where $f_* = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. □

## G. Convergence Analysis for MF-CA$^2$FL

*Proof of Theorem C.3.* Most of the proof for MF-CA$^2$FL follows the proof for CA$^2$FL. Denote $\widehat{\boldsymbol{v}}_t$ as the cached aggregated variable on the server, then we have $\mathbb{E}[\widehat{\boldsymbol{v}}_t] = \boldsymbol{v}_t$. Since $f$ is $L$-smooth, taking conditional expectation at time $t$, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1})] - f(\boldsymbol{x}_t) \\
&\leq \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] \\
&= \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \eta \widehat{\boldsymbol{v}}_t \rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|\widehat{\boldsymbol{v}}_t\|^2] \\
&= \underbrace{\mathbb{E}[\langle \nabla f(\boldsymbol{x}_t)), \eta \boldsymbol{v}_t \rangle]}_{I} + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}[\|\widehat{\boldsymbol{v}}_t\|^2]}_{II}.
\end{aligned} \tag{G.1}
$$

Note that term *I* is exactly the same as term *I* for CA$^2$FL. Hence we mainly show the proof for term *II*.

**Bounding *II***

$$
\begin{aligned}
II &= \frac{\eta^2 L}{2}\mathbb{E}[\|\widehat{\boldsymbol{v}}_t\|^2] = \frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\widehat{\Delta}^i_{t-\tau^i_t} - \widehat{\boldsymbol{h}}^i_t) + \frac{1}{N}\sum_{i=1}^N \widehat{\boldsymbol{h}}^i_t\bigg\|^2\bigg] \\
&= \frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\widehat{\Delta}^i_{t-\tau^i_t} - \widehat{\boldsymbol{h}}^i_t) + \frac{1}{N}\sum_{i=1}^N \widehat{\boldsymbol{h}}^i_t - \mathcal{Q}\bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg) \\
&\quad + \mathcal{Q}\bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg) - \bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg)\bigg\|^2\bigg] \\
&\leq \eta^2 L\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\widehat{\Delta}^i_{t-\tau^i_t} - \widehat{\boldsymbol{h}}^i_t) + \frac{1}{N}\sum_{i=1}^N \widehat{\boldsymbol{h}}^i_t - \mathcal{Q}\bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg)\bigg\|^2\bigg] \\
&\quad + \eta^2 L\mathbb{E}\bigg[\bigg\|\mathcal{Q}\bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg) - \bigg(\frac{1}{M}\sum_{i\in\mathcal{S}_t}(\Delta^i_{t-\tau^i_t} - \boldsymbol{h}^i_t) + \frac{1}{N}\sum_{i=1}^N \boldsymbol{h}^i_t\bigg)\bigg\|^2\bigg] \\
&\leq \eta^2 L(\gamma^2 + q^2)\mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}\Delta^i_{t-\tau^i_t} + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\sum_{i\in\mathcal{S}_t}\boldsymbol{h}^i_t + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\boldsymbol{h}^i_t\bigg\|^2\bigg].
\end{aligned} \tag{G.2}
$$

Therefore, by following the proof for Theorem 4.1 in Section F, we get the similar result as follows,

$$
\begin{aligned}
II &= \frac{\eta^2 L}{2}\mathbb{E}[\|\widehat{\boldsymbol{v}}_t\|^2] \\
&\leq \eta^2\eta_l^2 L(\gamma^2 + q^2)\frac{3K}{M}\sigma^2 + \eta^2\eta_l^2 L(\gamma^2 + q^2)\mathbb{E}\bigg[\bigg\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\bigg(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \bigg(\frac{1}{N} - \frac{1}{M}\bigg)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg) \\
&\quad + \frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\bigg\|^2\bigg].
\end{aligned} \tag{G.3}
$$

**Merging pieces.** For simplicity, we define $\mathbf{V}_t = \sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\big(\frac{1}{M}\nabla F_i(\boldsymbol{x}^i_{t-\tau^i_t,k}) + \big(\frac{1}{N} - \frac{1}{M}\big)\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})\big) +$

$\frac{1}{N}\sum_{i\notin\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}^i_{t-\zeta^i_t,k})$. Therefore, by merging pieces together, we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{t+1})]-f(\boldsymbol{x}_t)=I+II\\
&\leq-\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]-\frac{\eta\eta_l}{2K}\mathbb{E}[\|\mathbf{V}_t\|^2]+\bigg(3\eta\eta_l K+\frac{6\eta\eta_l K(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2+6K\sigma_g^2)\\
&\quad+3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]+\frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta^i_t})\|^2]\\
&\quad+\frac{3\eta\eta_l KL^2\tau_{\max}}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\eta^2L(\gamma^2+q^2)\bigg\{\frac{3K\eta_l^2}{M}\sigma^2+\eta_l^2\mathbb{E}[\|\mathbf{V}_t\|^2]\bigg\}\\
&=-\frac{\eta\eta_l K}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]+\bigg(3+\frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2L^2\eta_l^2(\sigma^2+6K\sigma_g^2)+\eta^2L(\gamma^2+q^2)\frac{3K\eta_l^2}{M}\sigma^2\\
&\quad+\frac{3\eta\eta_l KL^2\tau_{\max}}{M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]+\frac{3\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2M}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+\frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta^i_t}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&\quad+3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau^i_t})\|^2]+\frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta^i_t})\|^2]\\
&\quad-\bigg(\frac{\eta\eta_l}{2K}-\eta^2\eta_l^2L(\gamma^2+q^2)\bigg)\mathbb{E}[\|\mathbf{V}_t\|^2].
\end{aligned}
\tag{G.4}
$$

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

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{T+1})] - f(\boldsymbol{x}_1)\\
\leq & -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \left(3 + \frac{6(N-M)^2}{N^2}\right)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L(\gamma^2+q^2)\frac{3KT\eta_l^2}{M}\sigma^2\\
&+ \frac{3\eta\eta_l K L^2 \tau_{\max}}{M}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\tau_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg] + \frac{3\eta\eta_l K (N-M)^2\zeta_{\max}}{N^2 M}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\in\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&+ \frac{3\eta\eta_l K(N-M)\zeta_{\max}}{N^2}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{i\notin\mathcal{S}_t}\sum_{s=t-\zeta_t^i}^{t-1}\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|^2\bigg]\\
&+ 3\eta\eta_l K\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] + \frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]\\
&- \left(\frac{\eta\eta_l}{2K} - \eta^2\eta_l^2 L(\gamma^2+q^2)\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2]\\
\leq & -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \left(3 + \frac{6(N-M)^2}{N^2}\right)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L(\gamma^2+q^2)\frac{3KT\eta_l^2}{2M}\sigma^2\\
&+ \left(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2 (N-M)^2\zeta_{\max}^2}{N^2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2]\\
&+ 3\eta\eta_l K\sum_{t=1}^{T}\sum_{i\in\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\tau_t^i})\|^2] + \frac{6\eta\eta_l K(N-M)^2}{N^2}\sum_{t=1}^{T}\sum_{i\notin\mathcal{S}_t}30K^2L^2\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_{t-\zeta_t^i})\|^2]\\
&- \left(\frac{\eta\eta_l}{2K} - \eta^2\eta_l^2 L(\gamma^2+q^2)\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}_t\|^2],
\end{aligned}
\tag{G.5}
$$

while previously we obtained

$$
\mathbb{E}[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2] \leq 2\eta^2(\gamma^2+q^2)\frac{3K\eta_l^2}{M}\sigma^2 + 2\eta^2\eta_l^2(\gamma^2+q^2)\mathbb{E}[\|\mathbf{V}_t\|^2],
\tag{G.6}
$$

with the constraint of

$$
\begin{aligned}
&\eta^2\eta_l^2 L(\gamma^2+q^2) + 2\eta^2\eta_l^2(\gamma^2+q^2)\left(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\right) \leq \frac{\eta\eta_l}{2K}\\
\Rightarrow & \eta\eta_l(\gamma^2+q^2)KL + \eta^2\eta_l^2(\gamma^2+q^2)\left(12K^2L^2\tau_{\max}^2 + \frac{24K^2L^2(N-M)^2\zeta_{\max}^2}{N^2}\right) \leq 1\\
\Rightarrow & \eta\eta_l \leq \left(\sqrt{1 + \frac{48\tau_{\max}^2}{\gamma^2+q^2} + \frac{96(N-M)^2\zeta_{\max}^2}{N^2(\gamma^2+q^2)}} - 1\right)\left(12KL\tau_{\max}^2 + \frac{24KL(N-M)^2\zeta_{\max}^2}{N^2}\right)^{-1},
\end{aligned}
\tag{G.7}
$$

and

$$
\begin{aligned}
&\left(3\eta\eta_l K\tau_{\max} + \frac{6\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2}\right)(30K^2L^2\eta_l^2) \leq \frac{\eta\eta_l K}{4}\\
\Rightarrow & \eta_l \leq \left[\left(3\tau_{\max} + \frac{6(N-M)^2\zeta_{\max}}{N^2}\right)2\sqrt{30}KL\right]^{-1},
\end{aligned}
\tag{G.8}
$$

Thus we have

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}_{T+1})] - f(\boldsymbol{x}_1)\\
\leq & -\frac{\eta\eta_l K}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L(\gamma^2+q^2)\frac{3KT\eta_l^2}{M}\sigma^2\\
&+ \bigg(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{6\eta^2\eta_l^2 K(\gamma^2+q^2)T\sigma^2}{M}\\
&+ \bigg(3\eta\eta_l K\tau_{\max} + \frac{6\eta\eta_l K(N-M)^2\zeta_{\max}}{N^2}\bigg)(30K^2L^2\eta_l^2)\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]\\
\leq & -\frac{\eta\eta_l K}{4}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2 \eta_l^2(\sigma^2 + 6K\sigma_g^2) + \eta^2 L(\gamma^2+q^2)\frac{3KT\eta_l^2}{M}\sigma^2\\
&+ \bigg(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{6\eta^2\eta_l^2 K(\gamma^2+q^2)T\sigma^2}{M}. \qquad \text{(G.9)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{\eta\eta_l K}{4}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]\\
&\leq f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{T+1})] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5\eta\eta_l K^2 T L^2\eta_l^2(\sigma^2 + 6K\sigma_g^2)\\
&+ \eta^2 L(\gamma^2+q^2)\frac{3KT\eta_l^2}{M}\sigma^2 + \bigg(3\eta\eta_l K L^2\tau_{\max}^2 + \frac{6\eta\eta_l K L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{6\eta^2\eta_l^2 K(\gamma^2+q^2)T\sigma^2}{M},\\
\Rightarrow\quad &\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \frac{1}{\eta\eta_l KT}[f(\boldsymbol{x}_1) - \mathbb{E}[f(\boldsymbol{x}_{T+1})]] + \bigg(3 + \frac{6(N-M)^2}{N^2}\bigg)5KL^2\eta_l^2(\sigma^2 + 6K\sigma_g^2)\\
&+ \frac{3\eta\eta_l L(\gamma^2+q^2)}{M}\sigma^2 + \bigg(3L^2\tau_{\max}^2 + \frac{6L^2(N-M)^2\zeta_{\max}^2}{N^2}\bigg)\frac{6\eta^2\eta_l^2 K(\gamma^2+q^2)\sigma^2}{M}. \qquad \text{(G.10)}
\end{aligned}
$$

This concludes the proof. □

*Proof of Corollary C.4.* By choosing $\eta_l = \frac{1}{\sqrt{T}K}$ and $\eta = \sqrt{KM}$, then the convergence rate of MF-CA$^2$FL satisfies

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &= \mathcal{O}\bigg(\frac{f_0 - f_*}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{(\gamma^2+q^2)\sigma^2}{\sqrt{TKM}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2 + K\sigma_g^2}{TK}\bigg)\\
&+ \mathcal{O}\bigg(\frac{\tau_{\max}^2(\gamma^2+q^2)\sigma^2}{T}\bigg) + \mathcal{O}\bigg(\frac{\zeta_{\max}^2(N-M)^2(\gamma^2+q^2)\sigma^2}{TN^2}\bigg), \qquad \text{(G.11)}
\end{aligned}
$$

where $f_* = \mathrm{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. □

### G.1. Supporting Lemmas

**Lemma G.1.** *The global model difference $\Delta_t = \sum_{i\in\mathcal{S}_t}\Delta_t^i$ in partial participation cases satisfy*

$$\mathbb{E}[\|\Delta_t\|^2] = \frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2(N-M)}{NM(N-1)}[15NK^3L^3\eta_l^2(\sigma_l^2+6K\sigma_g^2) + 90NK^4L^2\eta_l^2 + 3NK^2\|\nabla f(\boldsymbol{x}_t)\|^2 + 3NK^2\sigma_g^2] + \frac{\eta_l^2(M-1)}{NM(N-1)}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2\bigg].$$

*Proof.* We have

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\bigg[\bigg\|\frac{1}{M}\sum_{i\in\mathcal{S}_t}\Delta_t^i\bigg\|^2\bigg]\\
&= \frac{1}{M^2}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\mathbb{I}\{i\in\mathcal{S}_t\}\Delta_t^i\bigg\|^2\bigg]\\
&= \frac{\eta_l^2}{M^2}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}[\mathbf{g}_{t,k}^i-\nabla F_i(\boldsymbol{x}_{t,k}^i)]\bigg\|^2 + \bigg\|\sum_{i=1}^N\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2\bigg]\\
&= \frac{\eta_l^2}{M^2}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\mathbb{P}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}[\mathbf{g}_{t,k}^i-\nabla F_i(\boldsymbol{x}_{t,k}^i)]\bigg\|^2 + \bigg\|\sum_{i=1}^N\mathbb{P}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2\bigg]\\
&= \frac{\eta_l^2}{MN}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\sum_{k=0}^{K-1}[\mathbf{g}_{t,k}^i-\nabla F_i(\boldsymbol{x}_{t,k}^i)]\bigg\|^2\bigg] + \frac{\eta_l^2}{M^2}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\mathbb{P}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2\bigg]\\
&\le \frac{K\eta_l^2}{M}\sigma_l^2 + \frac{\eta_l^2}{M^2}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^N\mathbb{P}\{i\in\mathcal{S}_t\}\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2\bigg],
\end{aligned}\tag{G.12}$$

where the fifth equation holds due to $\mathbb{P}\{i\in\mathcal{S}_t\} = \frac{M}{N}$. Note that we have

$$\begin{aligned}
\bigg\|\sum_{i=1}^N\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2 &= \sum_{i=1}^N\bigg\|\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2 + \sum_{i\neq j}\bigg\langle\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i), \sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}_{t,k}^j)\bigg\rangle\\
&= \sum_{i=1}^N N\bigg\|\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i)\bigg\|^2 - \frac{1}{2}\sum_{i\neq j}\bigg\|\sum_{k=0}^{K-1}\nabla F_i(\boldsymbol{x}_{t,k}^i) - \sum_{k=0}^{K-1}\nabla F_j(\boldsymbol{x}_{t,k}^j)\bigg\|^2,
\end{aligned}\tag{G.13}$$

where the second equation holds due to $\|\sum_{i=1}^N\boldsymbol{x}_i\|^2 = \sum_{i=1}^N N\|\boldsymbol{x}_i\|^2 - \frac{1}{2}\sum_{i\neq j}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$. By the sampling strategy (without replacement), we have $\mathbb{P}\{i \in \mathcal{S}_t\} = \frac{M}{N}$ and $\mathbb{P}\{i, j \in \mathcal{S}_t\} = \frac{M(M-1)}{N(N-1)}$, thus we have

$$
\begin{aligned}
&\Big\| \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{P}\{i \in \mathcal{S}_t\} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 \\
&= \sum_{i=1}^{N} \mathbb{P}\{i \in \mathcal{S}_t\} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 + \sum_{i \neq j} \mathbb{P}\{i, j \in \mathcal{S}_t\} \Big\langle \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i), \sum_{k=0}^{K-1} \nabla F_j(\boldsymbol{x}_{t,k}^j) \Big\rangle \\
&= \frac{M}{N} \sum_{i=1}^{N} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 + \frac{M(M-1)}{N(N-1)} \sum_{i \neq j} \Big\langle \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i), \sum_{k=0}^{K-1} \nabla F_j(\boldsymbol{x}_{t,k}^j) \Big\rangle \\
&= \frac{M^2}{N} \sum_{i=1}^{N} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 - \frac{M(M-1)}{2N(N-1)} \sum_{i \neq j} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) - \sum_{k=0}^{K-1} \nabla F_j(\boldsymbol{x}_{t,k}^j) \Big\|^2 \\
&= \frac{M(N-M)}{N(N-1)} \sum_{i=1}^{N} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 + \frac{M(M-1)}{N(N-1)} \Big\| \sum_{i=1}^{N} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2,
\end{aligned}
\tag{G.14}
$$

where the third equation holds due to $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{2}[\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2]$ and the last equation holds due to $\frac{1}{2} \sum_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \sum_{i=1}^{N} N \|\boldsymbol{x}_i\|^2 - \| \sum_{i=1}^{N} \boldsymbol{x}_i \|^2$. Therefore, for the last term in (G.12), we have

$$
\mathbb{E}[\|\Delta_t\|^2] = \frac{K\eta_l^2}{M} \sigma_l^2 + \frac{\eta_l^2 (N-M)}{NM(N-1)} \sum_{i=1}^{N} \mathbb{E}\Big[ \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 \Big] + \frac{\eta_l^2 (M-1)}{NM(N-1)} \mathbb{E}\Big[ \Big\| \sum_{i=1}^{N} \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 \Big]. \tag{G.15}
$$

The second term in (G.15) is bounded partially following (Reddi et al., 2021),

$$
\begin{aligned}
\sum_{i=1}^{N} \Big\| \sum_{k=0}^{K-1} \nabla F_i(\boldsymbol{x}_{t,k}^i) \Big\|^2 &= \sum_{i=1}^{N} \mathbb{E} \Big\| \sum_{k=0}^{K-1} [\nabla F_i(\boldsymbol{x}_{t,k}^i) - \nabla F_i(\boldsymbol{x}_t) + \nabla F_i(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)] \Big\|^2 \\
&\leq 3 \sum_{i=1}^{N} \mathbb{E} \Big\| \sum_{k=0}^{K-1} [\nabla F_i(\boldsymbol{x}_{t,k}^i) - \nabla F_i(\boldsymbol{x}_t)] \Big\|^2 + 3NK^2 \sigma_g^2 + 3NK^2 \|\nabla f(\boldsymbol{x}_t)\|^2 \\
&\leq 3KL^2 \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}[\|\boldsymbol{x}_{t,k}^i - \boldsymbol{x}_t\|^2] + 3NK^2 \sigma_g^2 + 3NK^2 \|\nabla f(\boldsymbol{x}_t)\|^2 \\
&\leq 15NK^3L^3\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + (90NK^4L^2\eta_l^2 + 3NK^2)\|\nabla f(\boldsymbol{x}_t)\|^2 + 3NK^2\sigma_g^2,
\end{aligned}
\tag{G.16}
$$

where the last inequality holds by applying Lemma G.2 (also follows from Reddi et al. (2021)). Substituting (G.16) into (G.15), this concludes the proof. □

**Lemma G.2.** *(This lemma directly 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*

$$
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{x}_{t,k}^i - \boldsymbol{x}_t\|^2] \leq 5K\eta_l^2(\sigma_l^2 + 6K\sigma_g^2) + 30K^2\eta_l^2 \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]. \tag{G.17}
$$

*Proof.* The proof of Lemma G.2 is exactly same as the proof of Lemma 3 in Reddi et al. (2021). □