PersA-FL: Personalized Asynchronous Federated Learning

Mohammad Taha Toghani[a], Soomin Lee[b], César A. Uribe[a]

[a]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA
[b]Yahoo! Research, Sunnyvale, California, USA

ARTICLE HISTORY
Compiled October 5, 2023

ABSTRACT
We study the personalized federated learning problem under asynchronous updates. In this problem, each client seeks to obtain a personalized model that simultaneously outperforms local and global models. We consider two optimization-based frameworks for personalization: (i) Model-Agnostic Meta-Learning (MAML) and (ii) Moreau Envelope (ME). MAML involves learning a joint model adapted for each client through fine-tuning, whereas ME requires a bi-level optimization problem with implicit gradients to enforce personalization via regularized losses. We focus on improving the scalability of personalized federated learning by removing the synchronous communication assumption. Moreover, we extend the studied function class by removing boundedness assumptions on the gradient norm. Our main technical contribution is a unified proof for asynchronous federated learning with bounded staleness that we apply to MAML and ME personalization frameworks. For the smooth and non-convex functions class, we show the convergence of our method to a first-order stationary point. We illustrate the performance of our method and its tolerance to staleness through experiments for classification tasks over heterogeneous datasets.

KEYWORDS
Federated Learning; Personalization; Asynchronous Communication; Heterogeneous Data; Distributed Optimization; Staleness.

1. Introduction

Federated Learning (FL) is designed to facilitate distributed training of machine learning models across devices by exploiting the data and computation power available to them [38]. A major benefit of FL is its ability to allow training models on data distributed across multiple devices without centralization. This is particularly beneficial in situations with limited sensitive data [32, 33] where clients are reluctant to share their private data. At the same time, it is known that training over a larger set of data points improves the quality of the obtained model [68]. In such scenarios, FL enjoys the power of collaborative learning without relocating the data from its original source [32]. Nevertheless, FL poses challenges such as data heterogeneity (statistical diversity among clients) [11, 17, 34, 42], fairness [14, 43], privacy [24, 30, 55, 69], unreliable communication [60, 64], and staleness [3, 44, 53, 70].

This work was partly done while MTT interning at Yahoo! Research. Part of this material is based upon work supported by the National Science Foundation under Grants #2211815 and #2213568. Corresponding Author's Email: mttoghani@rice.edu

The common underlying assumption that determines the superiority of FL to individual local training is that the data points of all clients are coming from the same distribution, i.e., homogeneous data across clients. Consequently, FL can improve the quality of empirical loss minimization when data available on each device is limited; otherwise, each client may obtain a proper model without collaboration or communication with others. Therefore, FL[1] results in a common global model with better generalization across clients [46] compared to individual training. In heterogeneous data setups where clients hold samples from non-identical data distributions, a common (global) model may perform poorly on the local data points of each client. For instance, consider the next word prediction task on a smart keyboard [28], where each client has a unique writing style or emphasis on the vocabulary domain. In this example, the corresponding mobile application is supposed to suggest a set of words that will likely be selected as the next word in the sentence. This scenario clearly states a case with a heterogeneous data setup with a limited sample on each client's device. Thus, if each client trains a model independently, without collaboration with the other clients, the model will likely perform poorly on the new data due to sample limitation. Hence, the question arises about what will occur if the clients hold data samples from similar (but not identical) distributions.

In FL with heterogeneous data, an ideal scenario is to learn a globally common model easily adaptable to local data on each client, i.e., model fusion. This approach is known as *Personalized Federated Learning* (PFL), which strives to exploit both the shared and unshared information from the data of all clients. A solution to the model fusion in PFL is to apply transfer learning [12, 73] (e.g., fine-tuning) on a jointly trained model under FL. Interestingly, the centralized version of this problem has been extensively studied in Meta-Learning [66] and Multi-Task Learning [50], where the goal is to obtain a meta (global) model that with (potentially) minimal adaptation performs well on multiple tasks. Particularly, Model-Agnostic Meta-Learning (MAML) [21, 56] proposes an optimization-based formulation that aims to find an initial meta-model with proper performance after applying one or a few steps of (stochastic) gradient descent. The key property of MAML is its ability to gauge fine-tuning during the learning process. Multiple studies have been conducted on the convergence and generalization of MAML [8, 16, 18, 19, 22, 31] for various problems and setups. Fallah et al. [17] suggest the MAML formulation as a potential solution for PFL, and propose Per-FedAvg algorithm for collaborative learning with MAML personalized cost function. Dinh et al. [13] present pFedMe algorithm for PFL via adopting a different formulation for personalization, namely Moreau Envelopes (ME). The proposed algorithm is a joint bi-level optimization problem with personalized parameters which are regularized to be close to the global model. We will elaborate on these two formulations (MAML & ME) in Section 2. Additionally, several recent works have approached PFL mainly through optimization-based [5, 10, 20, 23, 26, 27, 29, 45, 46, 64, 72], or structure-based [9, 59, 65] techniques.

Scalability to large-scale setups with potentially many clients is another major challenge for FL. The proposed algorithms in this scheme, mostly require synchronous communications between the server and clients [10, 13, 17, 23, 38, 43, 47]. Such constraints impose considerable delays on the learning progress, since increasing the concurrency in synchronous updates decreases the training speed and quality. For example, limited communication bandwidth, computation power, and communication failures incur large delays in the training process. In cross-device FL, devices are naturally prone to

[1]We refer to Federated Learning with no personalization as FL.

update and communicate models under less restrictive rules, whereas clients may apply updates in an asynchronous fashion, i.e., staleness. Hogwild! [52] is one of the first efforts to model asynchrony in distributed setup with delayed updates. Multiple works have studied asynchronous training under different setups and assumptions [1, 4, 6, 15, 44, 49, 53]. Specifically, some recent seminal works have studied the convergence of asynchronous SGD-based methods, and show their convergence under certain assumptions on maximum or average delay [2, 37, 48, 61].[2] In decentralized setups, Hadjicostis et al. [25] propose a consensus algorithm called running-sum, which is robust to message losses. Furthermore, Olshevsky et al. [54] present a more general framework with robustness to asynchrony, message losses, and delays for both consensus and optimization problems [60, 64]. More closely, FL under stale updates has been thoroughly studied in [3, 40, 51, 57, 65, 70]. Particularly, Tziotis et al. [65] studies the existence of stragglers in PFL via shared representations, i.e., system and data heterogeneity in structure-based personalization.

The main contribution of paper [51] is on the server algorithm, where this paper proposes a more secure and robust algorithm by aggregating a buffer of asynchronous updates within a secure channel prior to sending them to the server. Whereas, our work focuses on scalability and personalization via asynchronous communication and learning personalized models.

In this work, we study the PFL problem under asynchronous communications to improve training concurrency, performance, and efficiency. We propose the PersA-FL algorithm, a novel personalized & asynchronous method that jointly addresses the heterogeneity and staleness in FL. We develop a technique based on asynchronous updates to resolve the communication bottleneck imposed by synchronized learning in PFL, where we improve the training scalability and performance. To the best of our knowledge, this is the first study on the intersection of staleness and personalization through the lens of optimization-based techniques. We summarize our contributions as follows:

- Through the integration of two personalization formulations, MAML & ME, we propose PersA-FL, an algorithm that allows personalized federated learning under asynchronous communications between the server and clients. Our proposed method consists of two algorithms from the perspectives of the server and clients. We present the client algorithm under three different options for the local updates, each addressing a separate formulation, (A) FedAsync, (B) PersA-FL-MAML, and (C) PersA-FL-ME.
- We present a new convergence analysis for Asynchronous Federated Learning (FedAsync) under smooth non-convex cost functions by removing the boundedness assumption from the gradient norm. Our analysis assumes bounded variance of stochasticity and heterogeneity, and bounded maximum delay. Hence, we improve the existing theory by extending the result to a broader function class, i.e., unbounded gradient norm.
- We show the convergence rate of PersA-FL-MAML based on the maximum delay and personalization budget under the same assumptions as Fallah et al. [17].[3] We highlight the impact of batch size in the biased stochastic estimation of the full gradients for the MAML cost. We present the communication and sample

[2] Mishchenko et al. [48] studies the convergence of distributed optimization for homogeneous strongly convex and smooth functions with no assumptions on maximum delay, i.e., unbounded staleness.

[3] Besides the assumptions for FedAsync, seminal works [17, 22, 56] assume second-order Lipschitzness, bounded variance, and bounded gradient in the analysis of MAML cost functions.

complexity to find an ε first-order stationary point for the proposed algorithm.

- We prove the convergence of PersA-FL-ME with no boundedness assumption on the gradient norm. We discuss the connection of convergence rate to the gradient estimation error and level of personalization. Compared to [13], we show an explicit dependence of convergence rate to the estimation error. Moreover, we relax the heterogeneity assumption in [13] allowing bounded population diversity instead of uniformly bounded heterogeneity. We determine the communication and local inexact solver complexity to find an ε first-order stationary point for this method.
- We present numerical experiments evaluating our proposed algorithm on heterogeneous MNIST and CIFAR10 with unbalanced distributions across the clients. We illustrate the advantages of our method in terms of performance and scalability to varying delays in setups with heterogeneity.

Table 1 illustrates the properties of our proposed method and provides a comparison between our algorithm and underlying analysis with related seminal works. As shown in this table, building upon the results in [13, 17], we extend the capability of FL to staleness. Table 1 also contains the convergence results for our proposed algorithm, which we will discuss in more details in Section 4.

The main difference between our method and the works in [51, 70] mainly lies in the client algorithm, where we consider three options (A, B, and C) for updating the parameters locally. Option A is similar to the client algorithm in [51, 70], but we improve the theoretical convergence results by removing the assumption on bounded gradients for this setup. Option B and Option C, along with the server algorithm are novel methods for personalized asynchronous federated learning. Nguyen et al. [51] characterize the server algorithm with a secure and robust update aggregation and [63] enhances its theoretical properties. Study of secure aggregation on the server side remains as a future direction for this work.

The remainder of this paper is organized as follows. In Section 2, we introduce the PFL setup and discuss the asynchronous communication framework between the server and clients. In Section 3, we describe our algorithm, PersA-FL, for PFL under staleness. In Section 4, we state the convergence result for our proposed algorithm along with the underlying assumptions and technical lemmas. We present the numerical experiments in Section 5. We finally end by concluding remarks in Section 6.

2. Problem Setup & Background

In this section, we first present the formal problem setup for FL [47], as well as the personalization formulations in MAML [17] and ME [13]. Then, we discuss the underlying communication setting under asynchronous updates.

2.1. Federated Learning Problem Setup

We consider a set of n clients and one server, where each client $i \in [n]$ holds a private function $f_i : \mathbb{R}^d \to \mathbb{R}$, and the goal is to collaboratively obtain a model $w \in \mathbb{R}^d$ that

Table 1. A comparison of related federated learning methods with convergence guarantees for smooth non-convex functions. Parameters τ, α, ν, and b respectively denote the maximum delay, `MAML` personalization stepsize, `ME` inexact gradient estimation error, batch size.

Algorithm	**& Reference**	**Personalized Cost**	**Asynchronous Updates**	**Unbounded Gradient**	**Convergence Rate**
`FedAvg`	McMahan et al. [47]	✗	✗	-	No Analysis
	Yu et al. [71]	✗	✗	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
	Wang et al. [67]	✗	✗	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
`FedAsync`	Xie et al. [70]	✗	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
	This Work	✗	✓	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
`FedBuff`	Nguyen et al. [51]	✗	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
	Toghani and Uribe [63]	✗	✓	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
`Per-FedAvg`	Fallah et al. [17]	✓	✗	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\alpha^2}{b}\right)$
`pFedMe`	Dinh et al. [13]	✓	✗	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\lambda^2\left(\frac{1}{b}+\nu^2\right)}{(\lambda-L)^2}\right)$
`PersA-FL-MAML`	This Work	✓	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)+\mathcal{O}\left(\frac{\alpha^2}{b}\right)$
`PersA-FL-ME`	This Work	✓	✓	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)+\mathcal{O}\left(\frac{\lambda^2}{(\lambda-L)^2}\nu^2\right)$

minimizes the local cost functions on average, as follows:

$$\begin{aligned} \min_{w\in\mathbb{R}^d} f(w) &:= \frac{1}{n}\sum_{i=1}^{n} f_i(w), \\ \text{with} \quad f_i(w) &:= \mathbb{E}_{\Xi_i\sim p_i}[\ell_i(w,\Xi_i)], \end{aligned} \tag{1}$$

where $\ell_i : \mathbb{R}^d \times \mathcal{S}_i \to \mathbb{R}$ is a cost function that determines the prediction error of some model $w \in \mathbb{R}^d$ over a single data point $\xi_i \in \mathcal{S}_i$ on client i, where ξ_i is a realization of $\Xi_i \sim p_i$, i.e., p_i is the client i's data distribution over $\mathcal{S}_i$, for $i \in [n]$. In the above definition, $f_i(\cdot)$ is the local cost function of client i, and $f(\cdot)$ denotes the global cost function, i.e., average loss. For instance, in a supervised learning setup with $\mathcal{Z}_i := \mathcal{X}_i \times \mathcal{Y}_i$, we have $\ell_i(w, \xi_i)$ as the prediction cost of some learning model parameterized by w for sample $\xi_i = (x, y)$, where $x \in \mathcal{X}_i$ and $y \in \mathcal{Y}_i$. Let $\mathcal{D}_i$ be a data batch with samples independently drawn from the distribution p_i. Then, the unbiased stochastic

cost associated with data batch $\mathcal{D}_i$ can be denoted as follows:

$$\tilde{f}_i(w, \mathcal{D}_i) \coloneqq \frac{1}{|\mathcal{D}_i|} \sum_{\xi_i \in \mathcal{D}_i} \ell_i(w, \xi_i), \tag{2}$$

where for simplicity, we assume that the size of all batches is larger than b. Then, according to the above definition, we can immediately infer that

$$\begin{aligned} \mathbb{E}_{p_i}\left[\tilde{f}_i(w, \mathcal{D}_i)\right] &= f_i(w), \\ \mathbb{E}_{p_i}\left[\nabla \tilde{f}_i(w, \mathcal{D}_i)\right] &= \nabla f_i(w), \\ \mathbb{E}_{p_i}\left[\nabla^2 \tilde{f}_i(w, \mathcal{D}_i)\right] &= \nabla^2 f_i(w). \end{aligned} \tag{3}$$

Several works have been proposed to solve (1) as a union of local and global optimization steps. For instance, `FedAvg` [47] suggests an iterative algorithm wherein at each round $t \geq 0$, (i) server transmits its current parameter w^t to a subset of the clients, (ii) each selected client updates the parameter locally, by applying Q sequential rounds of stochastic gradient descent (`SGD`) with respect to its local cost function, then (iii) the selected clients send back their local parameter to the server, and finally, (iv) the server aggregates the so-called local parameters to obtain a new global parameter w^{t+1}. As a result, clients minimize the average loss in (1) with less communication cost, i.e., fewer global rounds. Note that the underlying assumption for methods such as `FedAvg` is the possibility of synchronized communications between the selected clients and the server. The left chart in Figure 1 represents the communication and update schedule for `FedAvg`. The performance of `FL`-based methods depends on the similarity of distributions $\mathcal{D}_i$, thus, cases with heterogeneous datasets slow down the convergence. Karimireddy et al. [34] and [11] the effect of heterogeneity in the convergence speed. A solution of (1) is a common model for all the clients; hence no adaptation or fusion to each client's data. Next, we elaborate on the personalization concept and discuss two alternative problem formulations for (1).

2.2. Personalized Federated Learning

In the previous section, we explained how a solution to (1) performs well when the data is homogeneous, and the goal is to obtain a shared model. On the one hand, using a single common model, with no adaptation to each client, does not necessarily lead to a proper performance when dealing with heterogeneous datasets. On the other hand, when the data distributions of different clients share some similarities, e.g., bounded variance in their heterogeneity, and the number of data points on each client is limited, joint training with fusion improves the performance compared to individual locally trained models or `FL`. Therefore, learning a shared model with little fine-tuning, e.g., a few steps of `SGD` with respect to the local cost, may result in a proper personalized model.

Fallah et al. [17] proposed `Per-FedAvg` algorithm, which modifies the training loss function by taking advantage of the fact that fine-tuning will occur after training. The `MAML` formulation assumes a limited computational budget for personalization (fine-tuning) at each client. It then offers to look for an initial (global) parameter that performs well after it is updated with one or a few steps of `SGD`. In other words, [17]

define the MAML loss function for PFL as follows:

$$\begin{aligned} \min_{w\in\mathbb{R}^d} F^{(b)}(w) &\coloneqq \frac{1}{n}\sum_{i=1}^{n} F_i^{(b)}(w), \\ \text{with} \quad F_i^{(b)}(w) &\coloneqq f_i(w-\alpha\nabla f_i(w)), \end{aligned} \tag{4}$$

where $\alpha \geq 0$ is the MAML personalization stepsize. Solving (4) yields a global (meta) model that can be used to create a personalized model by applying one step of gradient descent with respect to individual loss functions. The degree of fine-tuning determines the personalization budget, which often controls the trade-off between having a local (personalized) or generic model, i.e., exploiting the shared and local knowledge simultaneously. In Problem (4), stepsize α determines the personalization budget, where $\alpha = 0$ implies FL in Problem (1). See [18, 31, 64] for the study of multi-step MAML. In a nutshell, Per-FedAvg proposes to minimize $F^{(b)}(w)$ via a similar paradigm as FedAvg. Hence, each client i computes the personalized gradient of its MAML cost in (4), which can be written as follows:

$$\nabla F_i^{(b)}(w) = \left[I-\alpha\nabla^2 f_i(w)\right]\nabla f_i\left(w-\alpha\nabla f_i(w)\right), \tag{5}$$

where in Per-FedAvg, the authors propose to compute a biased estimation of (5) using stochastic gradients/Hessian. We will elaborate on the stochastic approximation in Section 3.

On a separate note, one of the major challenges in Per-FedAvg is the computation of second-order information such as Hessian for large-scale models (large d). However, as proposed by [17], one can skip the Hessian in the gradient formulation (FO-MAML) or approximate it with first-order information (HF-MAML) [16].

As an alternative option to MAML formulation in (4), Dinh et al. [13] suggest solving the following optimization problem:

$$\begin{aligned} \min_{w\in\mathbb{R}^d} F^{(c)}(w) &\coloneqq \frac{1}{n}\sum_{i=1}^{n} F_i^{(c)}(w), \\ \text{with} \quad F_i^{(c)}(w) &\coloneqq \min_{\theta_i\in\mathbb{R}^d}\left[f_i(\theta_i)+\frac{\lambda}{2}\|\theta_i-w\|^2\right], \end{aligned} \tag{6}$$

where each function $F_i^{(c)}(w)$ is a local cost of personalized parameter $\theta_i \in \mathbb{R}^d$ by using the Moreau Envelope as a regularized loss function, and parameter $\lambda \geq 0$ determines the degree of personalization. In this setup, $\lambda = 0$ is equivalent to local training with no collaboration and as $\lambda \to \infty$, the formulation in (6) converges to FL in (1) with no personalization which is similar to the case in (4) with $\alpha = 0$. For non-extreme values of λ, the clients jointly learn a global model w and personalized parameters θ_i, which are regularized to remain close to w. Note that the gradient of $F_i^{(c)}(w)$ can be written as follows (please check out Appendix C to see the proof):

$$\nabla F_i^{(c)}(w) = \lambda\left(w-\hat{\theta}_i(w)\right), \tag{7}$$

$$\text{with} \quad \hat{\theta}_i(w) \coloneqq \arg\min_{\theta_i\in\mathbb{R}^d}\left[f_i(\theta_i)+\frac{\lambda}{2}\|\theta_i-w\|^2\right], \tag{8}$$

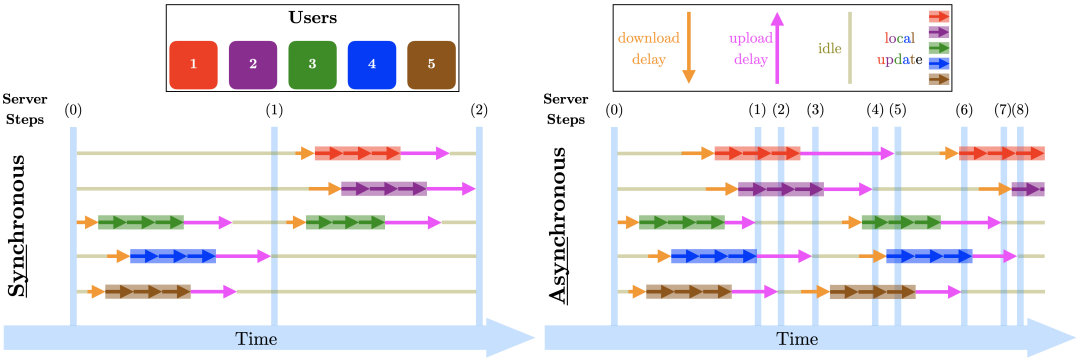


Figure 1. Communication and update schedule for synchronous and asynchronous aggregation: The demonstrated setup in this example contains $n = 5$ clients with $Q = 3$ local updates.

where for large λ, $\hat{\theta}_i(w)$ is the exact solution to an optimization problem. Therefore, solving (6) through a similar approach to `FedAvg` or `Per-FedAvg`, itself requires minimizing Problem (8) which is potentially intractable. Dinh et al. [13] propose a bi-level optimization algorithm called `pFedMe`, to minimize the optimization problem in (6) by alternating minimization over θ_i and w. The main idea behind `pFedMe` is to integrating the computation of an inexact solution to (8) inside an `FL`-type method. We will explain and use this inexact approximation in the presentation of our method (Option C) in Section 3.

2.3. Asynchronous vs Synchronous Schedule

So far, we have discussed the three different formulations for collaborative learning that we will consider in our method. As we described the `FedAvg` algorithm in Subsection 2.1, at each round t, the parameter w^t, which is the most recent version of the global parameter in the server, will be sent to a subset of the clients. Then, the server halts the training process until all selected clients receive this parameter, perform local updates, and transmit their updates back to the server. This synchronization procedure restricts the algorithm flow to the slowest client at each round. Nevertheless, asynchronous updates and communications can be described in this described framework.

Let us provide a comparison using the example in Figure 1 which illustrates the communication and update schedule for synchronous (left) & asynchronous (right) aggregations for $n = 5$ clients in `FL` with $Q = 3$ local updates. As shown in this Figure, for every update at the server-lever under synchronized updates (left figure), the server has to wait for all the selected clients. Nevertheless, these clients build their local updates based on the recent version of the server's parameter. On the contrary, in the asynchronous scenario (right figure), the server updates the global parameter once it receives a new update from some client. The main challenge for the asynchronous setup is the staleness between download and upload time from/to the server. We design `PersA-FL` based on the second communication scenario.

3. Algorithm: `FedAsync` & `PersA-FL`

In this section, by integrating the problem formulations in (1), (4), and (6) into a united format, we propose Algorithms 1 & 2 to solve these problems under three dif-

ferent update choices at the client-level. We present our method through two different perspectives, (i) server and (ii) client.

◇ **Server Algorithm:** Let us denote $w^0 \in \mathbb{R}^d$ as the initial parameter at the server, where the objective is to minimize the cost function in either (1), (4), or (6). Each client $i \in [n]$ may communicate with the server when the underlying connection is stable. Clients may request to download the server's parameters at any time, and the server will send the most recent model after receiving the request. All underlying delays for the communications between the server and clients are modeled as download and upload delays. We consider variable t as a counter for the updates at the server level. Algorithm 1 represents the server updates in `PersA-FL`. The server performs an iterative algorithm where at each round $t \geq 0$, remains on hold until receives an update $\Delta_{i_t} \in \mathbb{R}^d$ from some client $i_t \in [n]$. After receiving the update from client i_t, the server updates its parameter according to Step 4 of Algorithm 1, where $\beta \geq 0$ is the server stepsize.

Algorithm 1 [Personalized] Asynchronous Federated Learning (**Server**)

1: **input:** model w^0, $t = 0$, server stepsize β.
2: **repeat**
3: **if** the server receives an update Δ_{i_t} from some client $i_t \in [n]$ **then**
4: $w^{t+1} \leftarrow w^t - \beta \Delta_{i_t}$
5: $t \leftarrow t + 1$
6: **end if**
7: **until** not converge

Now, we are ready to present the client algorithm. Before starting, note that we drop the time index from the iterates of the client algorithm for clarity of exposition.

◇ **Client Algorithm**: Let us explain the operations of i-th client using the pseudo code in Algorithm 2. Client i repeats an iterative procedure which is composed of three phases, (i) downloading the most up-to-date model from the server as in Step 3, (ii) performing Q local updates starting from the parameters of the downloaded model with respect to the cost function of the underlying problem, (1), (4), or (6), as in Steps 5-13, and (iii) uploading the sum of updates on the server as in Step 15. Note that $\eta \geq 0$ is the local stepsize, a hyperparameter. The main idea for the local updates is to perform Q sequential `SGD` steps on the local cost. Below, we list our stochastic estimation for the full gradients of each loss function introduced in Section 2:

- **Option A:** This option intends to minimize (1). Therefore, for each client i at each local round q, we sample an independent data batch from p_i and compute an unbiased estimation of the loss as in (2).
- **Option B:** By performing this option, we aim to minimize the `MAML` cost function in (4). As we saw in Section 2, the full gradient can be computed according to (5). Following [17], we sample three data batches to compute a biased estimation of (5) as follows:

$$\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}_i'', \mathcal{D}_i', \mathcal{D}_i) = \left[I - \alpha \nabla^2 \tilde{f}_i(w, \mathcal{D}_i'')\right] \nabla \tilde{f}_i \left(w - \alpha \nabla \tilde{f}_i(w, \mathcal{D}_i'), \mathcal{D}_i\right). \tag{9}$$

  We will discuss the variance and bias of this estimator in Subsection 4.2
- **Option C:** Finally, we invoke this option to minimize the `ME` personalized loss in (6). As we mentioned earlier, the full gradient of this cost is (7), where for a fixed w, we may obtain $\hat{\theta}_i(w)$ by minimizing (8). Instead, following [13], we define the

Algorithm 2 [Personalized] Asynchronous Federated Learning (**Client** i)

1: **input:** number of local steps Q, local stepsize η, `MAML` stepsize α, `ME` regularization parameter λ, minimum batch size b, estimation error ν.
2: **repeat**
3: read w from the server ▷ download phase
4: $w_{i,0} \leftarrow w$
5: **for** $q = 0$ to $Q-1$ **do** ▷ local updates
6: sample a data batch $\mathcal{D}_{i,q}$ from distribution p_i ▽ 3 options:

▷ **Option A** (FedAsync)

7: $w_{i,q+1} \leftarrow w_{i,q} - \eta \nabla \tilde{f}_i(w_{i,q}, \mathcal{D}_{i,q})$

▷ **Option B** (`PersA-FL-MAML`)

8: sample two data batches $\mathcal{D}'_{i,q}, \mathcal{D}''_{i,q}$ from distribution p_i
9: $w_{i,q+1} \leftarrow w_{i,q} - \eta \left[I - \alpha \nabla^2 \tilde{f}_i(w_{i,q}, \mathcal{D}''_{i,q})\right] \nabla \tilde{f}_i \left(w_{i,q} - \alpha \nabla \tilde{f}_i(w_{i,q}, \mathcal{D}'_{i,q}), \mathcal{D}_{i,q}\right)$

▷ **Option C** (`PersA-FL-ME`)

10: $\tilde{h}_i(\theta_i, w_{i,q}, \mathcal{D}_{i,q}) \coloneqq \tilde{f}_i(\theta_i, \mathcal{D}_{i,q}) + \frac{\lambda}{2} \|\theta_i - w_{i,q}\|^2$
11: minimize $\tilde{h}_i(\theta_i, w_{i,q}, \mathcal{D}_{i,q})$ w.r.t. θ_i up to accuracy level ν to find $\tilde{\theta}_i(w_{i,q})$:

$$\left\| \nabla \tilde{h}_i \left(\tilde{\theta}_i(w_{i,q}), w_{i,q}, \mathcal{D}_{i,q} \right) \right\| \leq \nu$$

12: $w_{i,q+1} \leftarrow w_{i,q} - \eta\lambda(w_{i,q} - \tilde{\theta}_i(w_{i,q}))$

13: **end for**
14: $\Delta_i \leftarrow w_{i,0} - w_{i,Q}$
15: client i broadcasts Δ_i to the server ▷ upload phase
16: **until** not interrupted by the server

stochastic approximation $\tilde{h}_i(\theta_i, w, \mathcal{D}_i)$ as in Step 11, and minimize this function with respect to θ_i to obtain an approximate solution $\tilde{\theta}_i(w)$ where the gradient's norm is less than some threshold $\nu \geq 0$. Therefore, we approximate (7) with the following estimator:

$$\nabla \tilde{F}_i^{(c)}(w, \mathcal{D}_i) = \lambda \left(w - \tilde{\theta}_i(w) \right). \tag{10}$$

Let us denote the expectation of $\tilde{h}_i(.)$ as $h_i(.)$. Then, for $\lambda > L$, the expected function is $(\lambda+L)$-smooth and $(\lambda-L)$-strongly convex due to the properties of Moreau Envelopes [13]. Then according to the property of [7, 13], for some $\nu \leq 1$ (e.g., 10^{-5}), we can find $\tilde{\theta}_i(w)$ in $\mathcal{O}(\frac{\lambda+L}{\lambda-L} \log(\frac{1}{\nu}))$ iterations.

We will also discuss the properties of (10) in Subsection 4.3.

Next, we present the convergence result of our method for the three formulations.

4. Convergence Results

In this section, we introduce the technical theorems and lemmas to show the convergence of our method for the three described scenarios. First, we introduce the common assumptions we will use in our analysis for all the three choices of Algorithm 2. As

mentioned earlier, we require some additional assumptions to show the convergence of MAML, which we will introduce in Subsection 4.2. After stating the assumptions, we will present the convergence results.

Recall that the server updates its model at round t using the updates sent by client $i_t \in [n]$. We denote $\Omega(t)$ as the timestep of the round at which client i_t has received the server's parameters before applying its Q local updates. In other words, $(\Omega(t), t)$ denote the download and upload rounds for client i_t. Now, we introduce the assumption of maximum delay.

Assumption 1 (Bounded Staleness)**.** *For all server steps $t \geq 0$, the staleness or effective delay between the model version at the download step $\Omega(t)$ and upload step t is bounded by some constant τ, i.e.,*

$$\sup_{t \geq 0} |t - \Omega(t)| \leq \tau, \tag{11}$$

and the server receives updates uniformly, i.e., $i_t \sim \text{Uniform}([n])$.

The above assumption is standard in the analysis of asynchronous methods, specifically in heterogeneous settings [2, 3, 37, 51, 61, 70]. Assumption 1 guarantees that all clients remain active over the course of training. However, they have transient delays and perform updates with staleness.

Next, we present our only assumption on the function class, i.e., smooth non-convex.

Assumption 2 (Smoothness)**.** *For all clients $i \in [n]$, function $f_i : \mathbb{R}^d \to \mathbb{R}$ is bounded below, differentiable, and L-smooth, i.e., for all $w, u \in \mathbb{R}^d$,*

$$\|\nabla f_i(w) - \nabla f_i(u)\| \leq L\|w - u\| \tag{12}$$

$$f_i^\star \coloneqq \min_{w \in \mathbb{R}^d} f_i(w) > -\infty. \tag{13}$$

The smoothness assumption is conventional in the analysis of non-convex functions. We also assume boundedness from below, which is reasonable since the ultimate goal is to minimize the functions. We also denote $f^\star = \min_{i \in [n]} f_i^\star$, where according to this definition, we can immediately see that $f^\star \leq \min_{w \in \mathbb{R}^d} F^{(b)}(w)$ and $f^\star \leq \min_{w \in \mathbb{R}^d} F^{(c)}(w)$.

Now, we present our assumptions on bounded stochasticity and heterogeneity.

Assumption 3 (Bounded Variance)**.** *For all clients $i \in [n]$, the variance of a stochastic gradient $\nabla \ell_i(w, \xi_i)$ on a single data point $\xi_i \in \mathcal{S}_i$ is bounded, i.e., for all $w \in \mathbb{R}^d$*

$$\mathbb{E}_{\xi_i \sim p_i} \|\nabla \ell_i(w, \xi_i) - \nabla f_i(w)\|^2 \leq \sigma_g^2. \tag{14}$$

Assumption 3 is standard in the analysis of SGD-based methods and has been used in many relevant works [35–37, 51, 62, 64, 67]. Since we perform updates using data batches, we also need to show the stochastic variance for the sampled batches. Recall that for simplicity; we assumed that all batch sizes are larger than $b \geq 1$, thus, we have:

$$\mathbb{E}_{p_i} \left\| \nabla \tilde{f}_i(w, \mathcal{D}_i) - \nabla f_i(w) \right\|^2 \leq \frac{\sigma_g^2}{|\mathcal{D}_i|} \leq \sigma_a^2 \coloneqq \frac{\sigma_g^2}{b} \tag{15}$$

Next, we present the bounded heterogeneity assumption.

Assumption 4 (Bounded Population Diversity)**.** *For all $w \in \mathbb{R}^d$, the gradients of local functions $f_i(w)$ and the global function $f(w)$ satisfy the following property:*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w) - \nabla f(w)\|^2 \leq \gamma_g^2. \tag{16}$$

The above assumption measures the population diversity (heterogeneity) between the gradients. In heterogeneous settings, this bound indicates the similarity between different distributions. Fallah et al. [17] show connections between heterogeneity and the Wasserstein distance between the distributions under certain assumptions.

The above assumptions are sufficient to prove the convergence of our method (Algorithms 1 & 2) under Option A and Option C. Therefore, we present the convergence analyses starting from our results on `FedAsync`.

4.1. Asynchronous Federated Learning (Option A)

We now demonstrate the convergence rate of our method for the cost function in (1).

Theorem 1 (`FedAsync`)**.** *Let Assumptions 1-4 hold, $\beta = 1$, and $\eta = \frac{1}{Q\sqrt{LT}}$. Then, the following property holds for the joint iterates of Algorithms 1 & 2 under Option A on Problem (1): for any timestep $T \geq 160L(Q+7)(\tau+1)^3$ at the server*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f\left(w^t\right)\right\|^2 \leq \frac{4\sqrt{L}\left(f(w^0)-f^\star\right)}{\sqrt{T}} + \frac{8\sqrt{L}\left(\frac{\sigma_g^2}{b}+\gamma_g^2\right)}{\sqrt{T}} + \frac{80L(1+Q)(\tau^2+1)\left(\frac{\sigma_g^2}{b}+\gamma_g^2\right)}{T}.$$

The proof of Theorem 1 is provided in Appendix A. This theorem suggests a convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{Q\tau^2}{T}\right)$ for asynchronous federated learning `FedAsync`. Our analysis removes the unnecessary boundedness assumption on the gradient norm.

Remark 1. *Selecting $\beta = 1$ in Theorem 1, results in a sub-optimal first-order stationary rate for smooth non-convex cost functions. However, this is an arbitrary choice for the value of β and can be relaxed to any $\beta = \mathcal{O}(1)$ similar to [51].*

Next, we present the convergence of `PersA-FL-MAML` along with some technical lemmas borrowed from [17].

4.2. Personalized Asynchronous Federated Learning: Model-Agnostic Meta-Learning Setup (Option B)

As we discussed in Section 3, we require the second-order derivatives of the local functions to compute the gradients of the personalized costs in (4). Accordingly, we consider similar assumptions for the second-order derivatives as Assumptions 2-4.

Assumption 5 (Second-Order Properties)**.** *For all clients $i \in [n]$, the following properties hold for the Hessian of each $f_i : \mathbb{R}^d \to \mathbb{R}$, the variance of a stochastic Hessian*

$\nabla^2 \ell_i(w, \xi_i)$ *on a single data point* $\xi_i \in \mathcal{S}_i$*, and the global Hessian* $\nabla^2 f(w)$*: for all* $w, u \in \mathbb{R}^d$*,*

$$\left\|\nabla^2 f_i(w) - \nabla^2 f_i(u)\right\| \leq \rho \|w - u\|, \tag{17}$$

$$\mathbb{E}_{\xi_i \sim p_i} \left\|\nabla^2 \ell_i(w, \xi_i) - \nabla^2 f_i(w)\right\|^2 \leq \sigma_h^2, \tag{18}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left\|\nabla^2 f_i(w) - \nabla^2 f(w)\right\|^2 \leq \gamma_h^2. \tag{19}$$

Assumption 5 is conventional in the analysis of methods with access to second-order information [16, 17, 58, 64]. Finally, we adopt another assumption from [17, 19, 22] on the gradient norm to simplify the analysis for the `MAML` cost.

Assumption 6 (Bounded-Gradient)**.** *There exists a constant* G *such that for all clients* $i \in [n]$*, and any parameter* $w \in \mathbb{R}^d$*,*

$$\|\nabla f_i(w)\| \leq G. \tag{20}$$

To the best of our knowledge, seminal works on `MAML` loss mainly consider this assumption to simplify the properties of the personalized function. Note that we consider Assumptions 5-6 only in the analysis of `PersA-FL` (Algorithms 1 & 2) under Option B. Under Assumptions 2 and 6, the properties in (19) and (16) can be simply derived with $\gamma_h = 2L$ and $\gamma_g = 2G$ [17].

Before stating the convergence of `PersA-FL-MAML`, let us state some technical lemmas on the personalized `MAML` cost function.

Lemma 1 ([17], Lemma 4.2 - Smoothness: `MAML`)**.** *Let Assumptions 2 and 6 hold. Then,* $F_i^{(b)}$ *in* (4) *is* L_b*-smooth, i.e., for all clients* $i \in [n]$*, and any parameters* $w, u \in \mathbb{R}^d$*,*

$$\left\|\nabla F_i^{(b)}(w) - \nabla F_i^{(b)}(u)\right\| \leq L_b \|w - u\|, \tag{21}$$

where $L_b \coloneqq L(1+\alpha L)^2 + \alpha \rho G$*.*

Lemma 1 indicates that the personalized cost in (4) is also smooth. The smoothness parameter L_b depends on the personalization hyperparameter α. Increasing the value of α results in higher smoothness constant L_b. The smoothness property of `MAML` cost under multi-step personalization (instead of one) is shown in [64][Lemma 3].

Lemma 2 ([17], Lemma 4.3 - Bounded Variance: `MAML`)**.** *Let Assumptions 2, 3, 5, and 6 hold, and data batches* $\mathcal{D}, \mathcal{D}', \mathcal{D}''$ *be randomly sampled according to data distribution* p_i*. Then, the following properties hold for the stochastic personalized gradient* $\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D})$*:*

$$\left\|\mathbb{E}_{p_i}\left[\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D}) - \nabla F_i^{(b)}(w)\right]\right\| \leq \mu_b \coloneqq \frac{\alpha L(1+\alpha L)\sigma_g}{\sqrt{b}}, \tag{22}$$

$$\mathbb{E}_{p_i}\left\|\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D}) - \nabla F_i^{(b)}(w)\right\|^2 \leq \sigma_b^2, \tag{23}$$

for all $w \in \mathbb{R}^d$*, where* $\sigma_b^2 \coloneqq 3(1+\alpha L)^2 \sigma_g^2 \left[\frac{1}{b} + \frac{\alpha^2 L^2}{b}\right] + 3\alpha^2 G^2 \frac{\sigma_h^2}{b} + \frac{3\alpha^2 \sigma_g^2 \sigma_h^2}{b} \left[\frac{1}{b} + \frac{\alpha^2 L^2}{b}\right]$*.*

Lemma 2 highlights two important results. First, the stochastic gradient in (9) is a biased estimation of the full gradient 5. The biasness is controlled by two factors, personalization stepsize α, and batch size b.[4] Therefore, we obtain an unbiased estimation under no personalization, i.e., $\alpha = 0$. However, as we select a larger α, we require more samples to reduce the error imposed by biased gradient estimations. Second, similar to Assumption 3 on the cost; we have a tight variance based on α and b.

Lemma 3 ([17], Lemma 4.4 - Bounded Population Diversity: `MAML`)**.** *For all $w \in \mathbb{R}^d$, the gradients of local personalized functions $F_i^{(b)}(w)$ and the global function $F^{(b)}(w)$ satisfy the following property:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(b)}(w) - \nabla F^{(b)}(w)\right\|^2 \leq \gamma_b^2 \coloneqq 12(1+\alpha L)^2\left[1+\alpha^2 L^2\right]\gamma_g^2 + 12\alpha^2 G^2\gamma_h^2. \tag{24}$$

The above lemma determines the heterogeneity of the personalized gradients $\nabla F_i^{(b)}(w)$ based on the heterogeneity of gradient and Hessian. One can see the connection of this bound with $\mathcal{O}(\gamma_g^2) + \alpha^2\mathcal{O}(\gamma_h^2)$, whereby setting $\alpha = 0$, we recover the same heterogeneity in terms of $\mathcal{O}(\cdot)$ notion.

Lemma 4 (Bounded-Gradient: `MAML`)**.** *For all clients $i \in [n]$, and any parameter $w \in \mathbb{R}^d$,*

$$\left\|\nabla F_i^{(b)}(w)\right\| \leq G_b \coloneqq (1+\alpha L)G. \tag{25}$$

This lemma indicates that the bound on the norm of personalized gradients potentially increases under a larger personalization budget α.

Building upon the results in Lemmas 2-(25), we are now ready to present the convergence result for `PersA-FL-MAML`.

Theorem 2 (`PersA-FL-MAML`)**.** *Let Assumptions 1-6 hold, $\alpha \geq 0$, $\beta = 1$, and $\eta = \frac{1}{Q\sqrt{L_b T}}$. Then, the following property holds for the joint iterates of Algorithms 1 & 2 under Option B on Problem (4): for any timestep $T \geq 64L_b$ at the server*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \leq \frac{4\sqrt{L_b}\left(F^{(b)}(w^0) - f^\star\right)}{\sqrt{T}} + \frac{8\sqrt{L_b}\left(\sigma_b^2+\gamma_b^2\right)}{\sqrt{T}} + \frac{20\,QL_b\left(G_b^2+\sigma_b^2\right)\left(\tau^2+1\right)}{T} + \frac{4\,Q\alpha^2L^2(1+\alpha L)^2\sigma_g^2}{b}.$$

The proof of this theorem can be found in Appendix B. Theorem 2 shows a convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{\alpha^2\sigma_g^2}{b}\right)$ for `PersA-FL` algorithm under `MAML` setup. Now, let us compare this rate with the convergence rate of `FedAsync` and `Per-FedAvg`, as in Table 1. The last term in the above rate, i.e., $\mathcal{O}\left(\frac{\alpha^2\sigma_g^2}{b}\right)$ accounts for personalization with biased gradient estimation. Moreover, compared to `Per-FedAvg`, the second term of this rate is different, which accounts for the maximum delay in asynchronous

[4]It should be noted that the batch size in the upper bound of (22) refers to the size of $|\mathcal{D}'|$. Recall that we use this batch to approximate the inner gradient in 10.

updates.

To achieve the optimal complexity bound for the result in Theorem 2, we show how to choose the parameters T, b based on the desired accuracy ε in the following corollary.

Corollary 1 (`PersA-FL-MAML` ε-convergence)**.** *Suppose the conditions in Theorem 2 are satisfied. Algorithms 1 & 2 under Option B finds an ε first-order stationary solution for $F^{(b)}$ in (4) by setting $T = \mathcal{O}(\varepsilon^{-2})$ and $b = \mathcal{O}(\varepsilon^{-1})$ given a fixed personalization budget $\alpha \geq 0$.*

The result in Corollary 1 highlights the required communication and sample complexity for ε first-order stationary convergence. Moreover, note that the last expression in Theorem 2 can also be controlled through a combined stepsize α and batch size b. This result is consistent with intuition, i.e., more samples are required to obtain a higher degree of personalization.

Next, we will present the analysis of `PersA-FL-ME`.

4.3. Personalized Asynchronous Federated Learning: Moreau Envelope Setup (Option C)

In this subsection, we show three technical lemmas on the bounded variance of stochasticity and heterogeneity as well as smoothness for `ME` formulation (6) and then present the convergence rate of `PersA-FL` for this personalization framework. The proof of all results in this subsection is provided in Appendix C.

First, we present the smoothness property of `ME` loss.

Lemma 5 (Smoothness: ME)**.** *Let Assumption 2 holds and $\lambda \geq \kappa L$ for some $\kappa > 1$. Then, $F_i^{(c)}$ in (6) is L_c-smooth, where $L_c = \frac{\lambda}{\kappa - 1}$.*

According to Lemma 5, we limit our exploration to $\lambda > L$ which satisfies the smoothness constraint for the `ME` formulation. In fact, according to Appendix C, one can also see that originally, each $F_i^{(c)}(\cdot)$ is $\frac{\lambda L}{\lambda - L}$-smooth which is also smaller than $L_c = \frac{\lambda}{\kappa - 1}$. As we mentioned in Section 2, when $\lambda \to \infty$, `ME` framework converts to `FL`. The smoothness property in Lemma 5 is tight because, $L_c \to L$ if $\lambda \to \infty$.

Corollary 2 ([13], Proposition 1)**.** *If $\lambda \geq 2L$, then Lemma 5 implies that $F_i^{(c)}$ in (6) is λ-smooth.*

Lemma 6 (Bounded Variance: ME)**.** *Let Assumptions 2 and 3 hold, $\lambda \geq \kappa L$ (for some $\kappa > 1$), and the data batch $\mathcal{D}$ be randomly sampled according to data distribution p_i. Then, the following properties hold for the stochastic personalized gradient $\nabla \tilde{F}_i^{(c)}(w, \mathcal{D})$: for all $w \in \mathbb{R}^d$,*

$$\left\| \mathbb{E}_{p_i} \left[\nabla \tilde{F}_i^{(c)}(w, \mathcal{D}) - \nabla F_i^{(c)}(w) \right] \right\| \leq \mu_c \coloneqq \frac{\lambda}{\lambda - L} \nu, \tag{26}$$

$$\mathbb{E}_{p_i} \left\| \nabla \tilde{F}_i^{(c)}(w, \mathcal{D}) - \nabla F_i^{(c)}(w) \right\|^2 \leq \sigma_c^2 \coloneqq \frac{2\lambda^2}{(\lambda - L)^2} \left[\frac{\sigma_g^2}{b} + \nu^2 \right]. \tag{27}$$

This lemma is analogous to Lemma 2 in Subsection 4.2. In Lemma 6, we show an upper bound on the variance and bias of the stochastic gradient compared to the

full gradient. Note that when $\lambda \to \infty$, we know that $\hat{\theta}_i(w) \to w$. Therefore, by fixing $\tilde{\theta}_i(w) = w$, it is guaranteed that $\nu = 0$, thus our gradient estimation becomes unbiased and the variance similar to (15).

Lemma 7 (Bounded Population Diversity: ME)**.** *Let personalization hyperparameter $\lambda \geq 7L$. Then, for all $w \in \mathbb{R}^d$, the gradients of local personalized functions $F_i^{(c)}(w)$ and the global* ME *function $F^{(c)}(w)$ satisfy the following property:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\right\|^2 \leq \gamma_c^2 \coloneqq \frac{16\lambda^2}{\lambda^2 - 48L^2}\gamma_g^2. \tag{28}$$

Lemma 7 provides a bound on population diversity of ME as a factor of γ_g^2. Similar to what we explained so far, for $\lambda \to \infty$, the heterogeneity bound turns into γ_g^2.

Remark 2. *In the analysis for Theorem 3, we consider bounded population diversity as in Assumption 4, average bounded diversity. [13][Assumption 3] and [68][6.1.1 Assumptions and Preliminaries, (vii)] consider a slightly stronger version of this assumption, namely uniformly "bounded heterogeneity" which is defined as follows:*

$$\max_{i\in[n]}\sup_{w\in\mathbb{R}^d}\|\nabla f_i(w) - \nabla f(w)\|^2 \leq \gamma_g^2. \tag{29}$$

Under the modified assumption in (29)*, we can improve $\gamma_c^2 \coloneqq \frac{16\lambda^2}{\lambda^2 - 8L^2}\gamma_g^2$.*

Now, we present our convergence result of PersA-FL-ME under Assumption 1-4

Theorem 3 (PersA-FL-ME)**.** *Let Assumptions 1-4 hold, $\lambda \geq 7L$, $\beta = 1$, and $\eta = \frac{1}{Q\sqrt{L_c T}}$. Then, the following property holds for the joint iterates of Algorithms 1 & 2 under Option C on Problem* (6)*: for any timestep $T \geq 288L_c(Q+7)(\tau+1)^2$ at the server*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2 \leq \frac{4\sqrt{L_c}\left(F^{(c)}(w^t) - f^\star\right)}{\sqrt{T}} + \frac{8\sqrt{L_c}\left(\sigma_c^2 + \gamma_c^2\right)}{\sqrt{T}} + \frac{144L_c(1+Q)(\tau^2+1)\left(\sigma_c^2 + \gamma_g^2\right)}{T} + \frac{4Q\lambda^2\nu^2}{(\lambda - L)^2}.$$

Theorem 3 proposes a convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{\lambda^2\nu^2}{(\lambda-L)^2}\right)$ for PersA-FL under ME formulation. Again, under the exact same reasoning as Lemma 6, we know that $\nu = 0$ when $\lambda \to \infty$, thus the convergence rate simply reduces to FedAsync with no personalization. Moreover, let us compare the convergence result in Theorem 3 with the rate of pFedMe [13] in Table 1. By comparing the last terms in both rates, $\mathcal{O}\left(\frac{\lambda^2\nu^2}{(\lambda-L)^2}\right)$ and $\mathcal{O}\left(\frac{\lambda^2\left(\frac{1}{b}+\nu^2\right)}{(\lambda-L)^2}\right)$, one can see that the additional term $\frac{1}{b}$ in the convergence rate of pFedMe, implies that even under $\lambda \to \infty$ (i.e., no personalization), the last term does not vanish unless we select large data batches, i.e., $b = \mathcal{O}(\varepsilon^{-1})$. Therefore, from the personalization perspective, our analysis provides a tighter bound compared to pFedMe.

In the next corollary, we characterize a choice of T, ν in Theorem 3, given a desired accuracy level ε for our proposed algorithm in Option C.

Corollary 3 (PersA-FL-ME ε-convergence)**.** *Suppose the conditions in Theorem 3 are satisfied. Algorithms 1 & 2 under Option C finds an ε first-order stationary solution for $F^{(c)}$ in (6) by setting $T = \mathcal{O}(\varepsilon^{-2})$ and $\nu = \mathcal{O}(\varepsilon^{1/2})$ given a fixed personalization budget λ.*

Corollary 1 determines the communication complexity and precision of the approximate gradient estimator to achieve an ε first-order stationary convergence. This means that if we choose $\nu = \mathcal{O}(\varepsilon^{1/2})$, then the inexact optimization solver should compute the solution up to accuracy $O(\nu)$ of the surrogate optimization problem in order to achieve an ε-first order stationary solution. Also, we would like to highlight that this result implies no direct dependence on the batch size (b) for the convergence result of our algorithm with Option C(cf. [13]).

5. Numerical Experiments

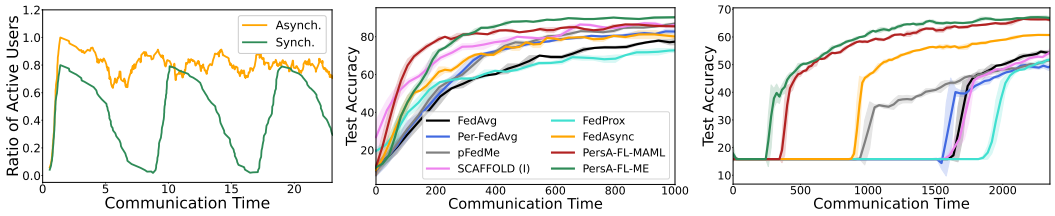
In this section, we evaluate the performance of our method in settings with delayed communications. We focus on the aspects of concurrency, speed-up, and accuracy.

Let us first start by explaining our simulation setup for communications with delays. We consider a set of n=30 different clients. Each of the clients has a set of random delays at the upload and download stage. The random delays are generated such that the average upload delay is 4 to 6 times higher than the average download delay. Moreover, we assume that the time for communication and aggregation is much larger than the time for local updates, thus we focus on the communication time. First, we show the number of active (not idle) users during the training process under asynchronous communications. The orange curve in Figure 2**(a)** shows the proportion of active users, which is up to 80% on average over time. We also plot the average proportion of users sampled in the synchronous updates in the same figure with green color. As Figure 2**(a)** demonstrates, the concurrency level for asynchronous methods is considerably higher than that of their synchronous counterparts.

We create extremely heterogeneous distributed data from MNIST [41] and CIFAR-10 [39] datasets for the clients, meaning that each client holds a different and skewed distribution of images from various classes. To build the heterogeneous data, we assign each client $i \in [n]$ samples from only c out of 10 classes of the data. Over the underlying communication setup and heterogeneous data setting, we compare the speed and accuracy of `FedAvg`, `Per-FedAvg`, `pFedMe`, `SCAFFOLD`[5], `FedAsync`, `PersA-FL-MAML`, `PersA-FL-ME`, where the first four methods are synchronous and the rest are asynchronous. For MNIST and CIFAR-10, we consider convolutional networks [39] followed by fully connected layers with pooling and dropout as well as cross-entropy loss. Details on the experimental setups can be found in AppendixD.

Figure 2 **(b)** and Figure 2 **(c)** compares the performance and convergence speed of our methods (`PersA-FL-MAML` & `PersA-FL-ME`) with the other five algorithms respectively on heterogeneous MNIST and CIFAR-10 datasets. *We would like to emphasize that in these two figures, each point on each curve represents the accuracy of the corresponding method after local fine-tuning with the same personalization budget as personalized personalized algorithm.* In other words, similar to the four personalized methods, `Per-FedAvg`, `pFedMe`, `PersA-FL-MAML`, and `PersA-FL-ME`, we consider same amount of fine-tuning budget for the three non-personalized methods. `FedAvg`,

[5]Scaffold algorithm has two options. Option I makes another pass over the local data to compute the gradient at the server model. Therefore, we consider SCAFFOLD (Option I), which is more stable in practice [34].



(a) Concurrency **(b) MNIST** **(c) CIFAR-10**

Figure 2. The impact of heterogeneity and communication delays on concurrency, convergence speed, and performance of multiple `FL`-based algorithms. The underlying setup of this experiment consists of n=30 clients, Q=10 local updates, and each client has a random upload and download delay at each round. **(a)** A comparison between the ratio of active users for synchronous and asynchronous updates over the course of training. **(b)** Comparison between the test accuracy of `FedAvg`, `Per-FedAvg`, `pFedMe`, `SCAFFOLD`, `FedProx`, `FedAsync`, `PersA-FL-MAML`, and `PersA-FL-ME` on MNIST data with heterogeneous distribution. **(c)** Test accuracy of the mentioned methods on CIFAR-10 data with synthetic heterogeneity within a limited fixed time.

`SCAFFOLD`, and `FedAsync`. As shown in Figure 2, our methods outperform the other methods within a fixed communication time. Moreover, the `ME` loss function results in a more stable and efficient performance compared to `MAML`.

6. Conclusion

This work studied the personalized federated learning problem for the heterogeneous data setting under asynchronous communications with the server. We considered the Model-Agnostic Meta-Learning (`MAML`) and Moreau Envelope (`ME`) formulations to account for personalization. We proposed the `PersA-FL` algorithm to solve this problem under stale updates. We showed the convergence rate of our method for smooth non-convex functions asynchronous federated learning, and personalized federated learning under the two personalization formulations, i.e., `MAML` and `ME`. Particularly, for asynchronous federated learning and personalized federated learning under Moreau Envelope costs, we presented a proof technique that does not require a boundedness assumption on the gradient norm. We finally show numerical results that illustrate the benefits of our proposed method in terms of accuracy and scalability. The studies of generalization and communication efficiency will be left to future research. Moreover, the extensions of our method to the buffered aggregation and decentralized setups remain for future studies.

References

[1] Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5451–5452. IEEE, 2012.

[2] Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pages 111–132. PMLR, 2020.

[3] Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

[4] Rotem Zamir Aviv, Ido Hakimi, Assaf Schuster, and Kfir Y Levy. Learning under delayed feedback: Implicitly adapting to gradient delays. *arXiv preprint arXiv:2106.12261*, 2021.

[5] El Houcine Bergou, Konstantin Burlachenko, Aritra Dutta, and Peter Richtárik. Personalized federated learning with communication compression. *arXiv preprint arXiv:2209.05148*, 2022.

[6] Dimitri Bertsekas. Distributed asynchronous policy iteration for sequential zero-sum games and minimax control. *arXiv preprint arXiv:2107.10406*, 2021.

[7] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[8] Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

[9] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

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

[11] Don Kurian Dennis and Virginia Smith. Heterogeneity for the win: Communication-efficient federated clustering, 2020.

[12] Dimitrios Dimitriadis, Kenichi Kumatani, Robert Gmyr, Yashesh Gaur, and Sefik Emre Eskimez. Federated transfer learning with dynamic gradient aggregation. *arXiv preprint arXiv:2008.02452*, 2020.

[13] Canh T Dinh, Nguyen H Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *arXiv preprint arXiv:2006.08848*, 2020.

[14] Siddharth Divi, Yi-Shan Lin, Habiba Farrukh, and Z Berkay Celik. New metrics to evaluate the performance and fairness of personalized federated learning. *arXiv preprint arXiv:2107.13173*, 2021.

[15] Hubert Eichner, Tomer Koren, Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning*, pages 1764–1773. PMLR, 2019.

[16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR, 2020.

[17] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020.

[18] Alireza Fallah, Kristian Georgiev, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of debiased model-agnostic meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Farzan Farnia, Amirhossein Reisizadeh, Ramtin Pedarsani, and Ali Jadbabaie. An optimal

transport approach to personalized federated learning. *arXiv preprint arXiv:2206.02468*, 2022.

[21] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[22] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

[23] Elnur Gasanov, Ahmed Khaled, Samuel Horváth, and Peter Richtárik. Flix: A simple and communication-efficient alternative to local methods in federated learning. *arXiv preprint arXiv:2111.11556*, 2021.

[24] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021.

[25] Christoforos N Hadjicostis, Nitin H Vaidya, and Alejandro D Domínguez-García. Robust distributed average consensus via exchange of running sums. *IEEE Transactions on Automatic Control*, 61(6):1492–1507, 2015.

[26] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33:2304–2315, 2020.

[27] Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint arXiv:2102.09743*, 2021.

[28] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[29] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7865–7873, 2021.

[30] Zhenqi Huang, Sayan Mitra, and Nitin Vaidya. Differentially private distributed optimization. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, pages 1–10, 2015.

[31] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Multi-step model-agnostic meta-learning: Convergence and improved algorithms. *arXiv preprint arXiv:2002.07836*, 2020.

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

[33] Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, pages 5201–5212. PMLR, 2021.

[34] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[35] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

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

[37] Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous sgd for distributed and federated learning. *arXiv preprint arXiv:2206.08307*, 2022.

[38] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha

Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[39] Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7):1–9, 2010.

[40] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797. IEEE, 2020.

[41] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[42] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[43] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[44] Yanan Li, Shusen Yang, Xuebin Ren, and Cong Zhao. Asynchronous federated learning with differential privacy for edge intelligence. *arXiv preprint arXiv:1912.07902*, 2019.

[45] Boxiang Lyu, Filip Hanzely, and Mladen Kolar. Personalized federated learning with multiple known clusters. *arXiv preprint arXiv:2204.13619*, 2022.

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

[47] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[48] Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake Woodworth. Asynchronous sgd beats minibatch sgd under arbitrary delays. *arXiv preprint arXiv:2206.07638*, 2022.

[49] Ioannis Mitliagkas, Ce Zhang, Stefan Hadjis, and Christopher Ré. Asynchrony begets momentum, with an application to deep learning. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 997–1004. IEEE, 2016.

[50] Matin Mortaheb, Cemil Vahapoglu, and Sennur Ulukus. Fedgradnorm: Personalized federated gradient-normalized multi-task learning. *arXiv preprint arXiv:2203.13663*, 2022.

[51] John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. Federated learning with buffered asynchronous aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 3581–3607. PMLR, 2022.

[52] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *arXiv preprint arXiv:1106.5730*, 2011.

[53] Kenta Niwa, Guoqiang Zhang, W Bastiaan Kleijn, Noboru Harada, Hiroshi Sawada, and Akinori Fujino. Asynchronous decentralized optimization with implicit stochastic variance reduction. In *International Conference on Machine Learning*, pages 8195–8204. PMLR, 2021.

[54] Alex Olshevsky, Ioannis Ch Paschalidis, and Artin Spiridonoff. Fully asynchronous push-sum with growing intercommunication intervals. In *2018 Annual American Control Conference (ACC)*, pages 591–596. IEEE, 2018.

[55] Karthik Prasad, Sayan Ghosh, Graham Cormode, Ilya Mironov, Ashkan Yousefpour, and Pierre Stock. Reconciling security and communication efficiency in federated learning. *arXiv preprint arXiv:2207.12779*, 2022.

[56] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[57] Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statis-

tical accuracy and system heterogeneity. *IEEE Journal on Selected Areas in Information Theory*, 2022.

[58] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

[59] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR, 2021.

[60] Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch Paschalidis. Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions. *Journal of machine learning research*, 21(58), 2020.

[61] Sebastian Stich, Amirkeivan Mohtashami, and Martin Jaggi. Critical parameters for scalable distributed learning with large batches and asynchronous updates. In *International Conference on Artificial Intelligence and Statistics*, pages 4042–4050. PMLR, 2021.

[62] Sebastian Urban Stich. Local sgd converges fast and communicates little. In *ICLR 2019-International Conference on Learning Representations*, number CONF, 2019.

[63] Mohammad Taha Toghani and César A Uribe. Unbounded gradients in federated learning with buffered asynchronous aggregation. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2022.

[64] Mohammad Taha Toghani, Soomin Lee, and César A. Uribe. Pars-push: Personalized, asynchronous and robust decentralized optimization. *IEEE Control Systems Letters*, 7: 361–366, 2023. .

[65] Isidoros Tziotis, Zebang Shen, Ramtin Pedarsani, Hamed Hassani, and Aryan Mokhtari. Straggler-resilient personalized federated learning. *arXiv preprint arXiv:2206.02078*, 2022.

[66] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[67] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[68] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[69] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[70] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

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

[72] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

[73] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

Appendix A. Asynchronous Federated Learning

Proof of Theorem 1. First, we present a set of useful inequalities we will use in the proof. For any set of m vectors $\{w_i\}_{i=1}^m$ such that $w_i \in \mathbb{R}^d$, and a constant $\alpha > 0$, the following properties hold: for all $i, j \in [m]$:

$$\|w_i + w_j\|^2 \leq (1+\alpha)\|w_i\|^2 + (1+\alpha^{-1})\|w_j\|^2, \tag{A1a}$$
$$\|w_i + w_j\| \leq \|w_i\| + \|w_j\|, \tag{A1b}$$
$$2\langle w_i, w_j\rangle \leq \alpha\|w_i\|^2 + \alpha^{-1}\|w_j\|^2, \tag{A1c}$$
$$\left\|\sum_{i=1}^m w_i\right\|^2 \leq m\left(\sum_{i=1}^m \|w_i\|^2\right). \tag{A1d}$$

Now, let us rewrite the update rule of the joint iterates in Algorithms 1 & 2 Option A at time t as follows:

- Client update:

$$w_{i,0}^t = w^t, \tag{A2}$$
$$w_{i,q+1}^t = w_{i,q}^t - \eta\nabla\tilde{f}_i(w_{i,q}^t, \mathcal{D}_{i,q}^t), \tag{A3}$$

- Server update:

$$w^{t+1} = w^t - \beta\Delta_{i_t} = w^t - \eta\beta\sum_{q=0}^{Q-1}\nabla\tilde{f}_{i_t}\left(w_{i_t,q}^{\Omega(t)}, \mathcal{D}_{i,q}^{\Omega(t)}\right). \tag{A4}$$

For simplicity, we denote $\tilde{\nabla} f_i(w) = \nabla\tilde{f}_i(w, \mathcal{D}_i)$. Therefore, at round t, the server updates its parameter by receiving Δ_{i_t} from some client $i_t \in [n]$, as follows:

$$w^{t+1} = w^t - \eta\beta\sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right). \tag{A5}$$

Moreover, Due to Assumption 2, we can infer that f is L-smooth, thus

$$f\left(w^{t+1}\right) \overset{(12)}{\leq} f(w^t) - \eta\beta\underbrace{\left\langle\nabla f(w^t), \sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\rangle}_{=:S_{a_1}} + \frac{L\eta^2\beta^2}{2}\underbrace{\left\|\sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\|^2}_{=:S_{a_2}} \tag{A6}$$

First, we provide a lower bound on term S_{a_1} in (A6). Prior to show the bound, let us denote $\tilde{g}_i^t = \sum_{q=0}^{Q-1}\tilde{\nabla} f_i\left(w_{i,q}^{\Omega(t)}\right)$, $\tilde{g}^t = \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^t$, $g_i^t = \sum_{q=0}^{Q-1}\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)$, and $g^t = \frac{1}{n}\sum_{i=1}^n g_i^t$.

Therefore,

$$\mathbb{E}\left[S_{a_1}\right] = \mathbb{E}\left[\mathbb{E}_{i_t}\left\langle \nabla f(w^t), \tilde{g}_{i_t}^t \right\rangle\right] \tag{A7}$$

$$= \mathbb{E}\left[\left\langle \nabla f(w^t), \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^t \right\rangle\right] \tag{A8}$$

$$= \mathbb{E}\left\langle \nabla f(w^t), \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p_i}\left[\tilde{g}_i^t\right] \right\rangle = \mathbb{E}\left\langle \nabla f(w^t), \frac{1}{n}\sum_{i=1}^n g_i^t \right\rangle \tag{A9}$$

$$= Q\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 + \mathbb{E}\left\langle \nabla f(w^t), g^t - Q\nabla f(w^t)\right\rangle \tag{A10}$$

$$\overset{\text{(A1c)}}{\geq} Q\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|g^t - Q\nabla f(w^t)\right\|^2 \tag{A11}$$

$$= \frac{2Q-1}{2}\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|g^t - Q\nabla f(w^t)\right\|^2. \tag{A12}$$

Moreover, the following holds for S_{a_2} in (A6):

$$\mathbb{E}_{i_t}\left[S_{a_2}\right] = \mathbb{E}_{i_t}\left\|\sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|\tilde{g}_i^t\right\|^2. \tag{A13}$$

Now, according to (A6), (A12), and (A13), we have:

$$\mathbb{E}f\left(w^{t+1}\right) \leq \mathbb{E}f(w^t) - \frac{\eta\beta(2Q-1)}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{A14}$$

$$+ \frac{\eta\beta}{2}\mathbb{E}\underbrace{\left\|g^t - Q\nabla f(w^t)\right\|^2}_{=:S_{a_3}} + \frac{L\eta^2\beta^2}{2n}\mathbb{E}\underbrace{\left[\sum_{i=1}^n \left\|\tilde{g}_i^t\right\|^2\right]}_{=:S_{a_4}}, \tag{A15}$$

where we bound S_{a_3} and S_{a_4} as follows:

$$S_{a_3} = \left\|\frac{1}{n}\sum_{i=1}^n \left(g_i^t - Q\nabla f_i(w^t)\right)\right\|^2 \overset{\text{(A1d)}}{\leq} \frac{1}{n}\sum_{i=1}^n \left\|g_i^t - Q\nabla f_i(w^t)\right\|^2 \tag{A16}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\|\sum_{q=0}^{Q-1}\nabla f_i\left(w_{i,q}^{\Omega(t)}\right) - Q\nabla f_i(w^t)\right\|^2 \tag{A17}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\|\sum_{q=0}^{Q-1}\left[\nabla f_i\left(w_{i,q}^{\Omega(t)}\right) - \nabla f_i(w^t)\right]\right\|^2 \tag{A18}$$

$$\overset{\text{(A1d)}}{\leq} \frac{Q}{n}\sum_{i=1}^n\sum_{q=0}^{Q-1}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right) - \nabla f_i(w^t)\right\|^2, \tag{A19}$$

$$S_{a_4} = \sum_{i=1}^{n} \Big\| \sum_{q=0}^{Q-1} \tilde{\nabla} f_i \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \tag{A20}$$

$$\overset{\text{(A1d)}}{\leq} Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \Big\| \tilde{\nabla} f_i \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \tag{A21}$$

$$= Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \Big\| \tilde{\nabla} f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) + \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w^t \right) + \nabla f_i \left(w^t \right) - \nabla f \left(w^t \right) + \nabla f \left(w^t \right) \Big\|^2 \tag{A22}$$

$$\overset{\text{(A1d)}}{\leq} 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \left[\Big\| \tilde{\nabla} f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 + \Big\| \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w^t \right) \Big\|^2 + \Big\| \nabla f_i \left(w^t \right) - \nabla f \left(w^t \right) \Big\|^2 + \Big\| \nabla f \left(w^t \right) \Big\|^2 \right] \Rightarrow \tag{A23}$$

$$\mathbb{E}[S_{a_4}] \overset{\text{(A23)}}{\leq} 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E}_{p_i} \left[\Big\| \tilde{\nabla} f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \right] \tag{A24}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i \left(w^t \right) \Big\|^2 \tag{A25}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla f_i \left(w^t \right) - \nabla f \left(w^t \right) \Big\|^2 \tag{A26}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla f \left(w^t \right) \Big\|^2 \tag{A27}$$

$$\overset{\text{(15),(16)}}{\leq} 4nQ^2 \left[\sigma_a^2 + \gamma_g^2 + \mathbb{E} \Big\| \nabla f \left(w^t \right) \Big\|^2 \right] \tag{A28}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla f_i \left(w_{i,q}^{\Omega(t)} \right) - \nabla f_i (w^t) \Big\|^2 . \tag{A29}$$

Therefore, due to (A14)-(A19) and (A28)-(A29), we have

$$\mathbb{E}f\left(w^{t+1}\right) \leq \mathbb{E}f(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L\beta^2 Q^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{A30}$$

$$+ \left[\frac{\eta\beta Q}{2n} + \frac{2\eta^2\beta^2 QL}{n}\right]\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right) - \nabla f_i(w^t)\right\|^2 \tag{A31}$$

$$+ 2\eta^2 L\beta^2 Q^2\sigma_a^2 + 2\eta^2 L\beta^2 Q^2\gamma_g^2 \tag{A32}$$

$$\overset{(12)}{\leq} \mathbb{E}f(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L\beta^2 Q^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{A33}$$

$$+ \frac{\eta\beta QL^2\left(1+4\eta\beta L\right)}{2n}\mathbb{E}\underbrace{\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|w_{i,q}^{\Omega(t)} - w^t\right\|^2}_{=:S_{a_5}} \tag{A34}$$

$$+ 2\eta^2 L\beta^2 Q^2\sigma_a^2 + 2\eta^2 L\beta^2 Q^2\gamma_g^2. \tag{A35}$$

Thus, it is sufficient to bound the following expression in S_{a_5}:

$$\left\|w^t - w_{i,q}^{\Omega(t)}\right\|^2 \tag{A36}$$

$$= \left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1} - w^s\right) + w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{A37}$$

$$\overset{(A1a)}{\leq} \left(1+\frac{1}{\beta^2}\right)\left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1} - w^s\right)\right\|^2 + \left(1+\beta^2\right)\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{A38}$$

$$\overset{(A1d)}{\leq} (t-\Omega(t))\left(1+\frac{1}{\beta^2}\right)\left[\sum_{s=\Omega(t)}^{t-1}\left\|w^{s+1} - w^s\right\|^2\right] + \left(1+\beta^2\right)\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{A39}$$

$$\overset{(11)}{\leq} \tau\left(1+\frac{1}{\beta^2}\right)\underbrace{\left[\sum_{s=t-\tau}^{t-1}\left\|w^{s+1} - w^s\right\|^2\right]}_{=:S_{a_7}} + \left(1+\beta^2\right)\underbrace{\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2}_{=:S_{a_6}}. \tag{A40}$$

Now, we show a bound on the evolution of local updates at an arbitrary round

$s \geq 0$, i.e., the distance between $w_{i,q}^s$ and w^s:

$$\mathbb{E}\left\|w_{i,q}^s - w^s\right\|^2 = \mathbb{E}\left\|w_{i,q-1}^s - \eta\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) - w^s\right\|^2 \tag{A41}$$

$$\begin{aligned} &= \mathbb{E}\Big\|w_{i,q-1}^s - w^s - \eta\nabla f\left(w^s\right) \\ &\quad - \eta\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) + \eta\nabla f_i\left(w_{i,q-1}^s\right) \\ &\quad - \eta\nabla f_i\left(w_{i,q-1}^s\right) + \eta\nabla f_i\left(w^s\right) \\ &\quad - \eta\nabla f_i\left(w^s\right) + \eta\nabla f\left(w^s\right)\Big\|^2 \end{aligned} \tag{A42}$$

$$\begin{aligned} &\overset{\text{(A1a)}}{\leq} \left(1+\frac{1}{2Q}\right)\mathbb{E}\left\|w_{i,q-1}^s - w^s\right\|^2 \\ &+ \quad 4(1+2Q)\eta^2\mathbb{E}\Bigg[\left\|\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) - \nabla f_i\left(w_{i,q-1}^s\right)\right\|^2 \\ &\qquad + \left\|\nabla f_i\left(w_{i,q-1}^s\right) - \nabla f_i\left(w^s\right)\right\|^2 \\ &\qquad + \left\|\nabla f_i\left(w^s\right) - \nabla f\left(w^s\right)\right\|^2 \\ &\qquad + \left\|\nabla f\left(w^s\right)\right\|^2\Bigg] \end{aligned} \tag{A43, A44}$$

$$\begin{aligned} &\overset{(12),(15)}{\leq} \left(1+\frac{1}{2Q}\right)\mathbb{E}\left\|w_{i,q-1}^s - w^s\right\|^2 \\ &+ \quad 4(1+2Q)\eta^2\Bigg[\sigma_a^2 + L^2\,\mathbb{E}\left\|w_{i,q-1}^s - w^s\right\|^2 \\ &\qquad + \mathbb{E}\left\|\nabla f_i\left(w^s\right) - \nabla f\left(w^s\right)\right\|^2 + \mathbb{E}\left\|\nabla f\left(w^s\right)\right\|^2\Bigg]. \end{aligned} \tag{A45, A46}$$

Note that we can select stepsize $\eta \leq \frac{1}{4L(Q+1)}$ such that

$$\eta^2 \leq \frac{1}{16L^2(Q+1)^2} \leq \frac{1}{8L^2Q(2Q+1)} \Rightarrow 4(1+2Q)\eta^2L^2 \leq \frac{1}{2Q}, \tag{A47}$$

therefore, due to (A41)-(A47), we have:

$$\underbrace{\mathbb{E}\left\|w_{i,q}^{s}-w^{s}\right\|^{2}}_{:=P_{i,q}^{s}}\leq\underbrace{\left(1+\frac{1}{Q}\right)\mathbb{E}\left\|w_{i,q-1}^{s}-w^{s}\right\|^{2}}_{:=P_{i,q-1}^{s}} \tag{A48}$$

$$+\underbrace{4(1+2Q)\eta^{2}\left[\sigma_{a}^{2}+\mathbb{E}\left\|\nabla f_{i}\left(w^{s}\right)-\nabla f\left(w^{s}\right)\right\|^{2}+\mathbb{E}\left\|\nabla f\left(w^{s}\right)\right\|^{2}\right]}_{:=R_{i}^{s}}\Rightarrow \tag{A49}$$

$$P_{i,q}^{s}\leq\left(1+\frac{1}{Q}\right)P_{i,q-1}^{s}+R_{i}^{s} \tag{A50}$$

$$=R_{i}^{s}\sum_{k=0}^{q-1}\left(1+\frac{1}{Q}\right)^{k}\leq R_{i}^{s}\sum_{k=0}^{Q-1}\left(1+\frac{1}{Q}\right)^{k} \tag{A51}$$

$$=R_{i}^{s}\frac{\left(1+\frac{1}{Q}\right)^{Q}-1}{\left(1+\frac{1}{Q}\right)-1}=R_{i}^{s}Q\left[\left(1+\frac{1}{Q}\right)^{Q}-1\right]\leq R_{i}^{s}Q(e-1)\leq 2R_{i}^{s}Q\Rightarrow \tag{A52}$$

$$\mathbb{E}\left\|w_{i,q}^{s}-w^{s}\right\|^{2}\leq 8Q(1+2Q)\eta^{2}\left[\sigma_{a}^{2}+\mathbb{E}\left\|\nabla f_{i}\left(w^{s}\right)-\nabla f\left(w^{s}\right)\right\|^{2}+\mathbb{E}\left\|\nabla f\left(w^{s}\right)\right\|^{2}\right], \tag{A53}$$

for all $q\in[Q]$ and $s\geq 0$. We now will use (A41)-(A53) to provide a bound on the expression in S_{a_7}. Again, note that according to Algorithms 1 & 2, we have:

$$w^{s+1}=w^{s}-\beta\left(w_{i_s,0}^{\Omega(s)}-w_{i_s,Q}^{\Omega(s)}\right)\Rightarrow \tag{A54}$$

$$\mathbb{E}\left\|w^{s+1}-w^{s}\right\|^{2}\leq\beta^{2}\,\mathbb{E}\left\|w_{i_s,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2} \tag{A55}$$

$$=\beta^{2}\,\mathbb{E}\left[\mathbb{E}_{i_s}\left\|w_{i_s,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2}\right] \tag{A56}$$

$$=\frac{\beta^{2}}{n}\sum_{j=1}^{n}\mathbb{E}\left\|w_{j,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2} \tag{A57}$$

$$\leq 8Q(1+2Q)\eta^{2}\beta^{2}\left[\sigma_{a}^{2}+\gamma_{g}^{2}+\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^{2}\right]. \tag{A58}$$

Let $\phi = 8\eta^2 Q^2(1+2Q)(1+\beta^2)$, then according to (A36)-(A58)

$$\frac{1}{n\phi}\mathbb{E}[S_{a_5}] \leq \tau \left[\sum_{s=t-\tau}^{t-1} \left\|w^{s+1} - w^s\right\|^2\right] + \frac{1}{nQ}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2. \tag{A59}$$

$$\leq \tau^2\sigma_a^2 + \tau^2\gamma_g^2 + \tau\sum_{s=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2 \tag{A60}$$

$$+ \sigma_a^2 + \gamma_g^2 + \mathbb{E}\left\|\nabla f\left(w^{\Omega(t)}\right)\right\|^2 \tag{A61}$$

$$\leq (\tau^2+1)\left[\sigma_a^2 + \gamma_g^2\right] + \mathbb{E}\left\|\nabla f\left(w^{\Omega(t)}\right)\right\|^2 + \tau\sum_{s=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2 \tag{A62}$$

$$\leq (\tau^2+1)\left[\sigma_a^2 + \gamma_g^2\right] + \tau\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2. \tag{A63}$$

Thus, by combining (A30)-(A63), we have the following inequality:

$$\mathbb{E}f\left(w^{t+1}\right) \leq \mathbb{E}f(w^t) - \eta\beta\left[\frac{2Q-1}{2} - 2\eta\beta LQ^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{A64}$$

$$+ 4\eta^3\beta L^2Q^3(1+2Q)(1+\beta^2)(1+4\eta\beta L)\,\tau\left[\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2\right] \tag{A65}$$

$$+ 4\eta^3\beta L^2Q^3(1+2Q)(\tau^2+1)(1+\beta^2)(1+4\eta\beta L)\left(\sigma_a^2 + \gamma_g^2\right) \tag{A66}$$

$$+ 2\eta^2\beta^2 LQ^2\left(\sigma_a^2 + \gamma_g^2\right), \tag{A67}$$

where by rearranging, we obtain the following inequality:

$$(1 - 4\eta\beta LQ)\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{A68}$$

$$- 8\eta^2L^2Q^2(1+2Q)(1+\beta^2)(1+4\eta\beta L)\,\tau\left[\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2\right] \tag{A69}$$

$$\leq \frac{2\left[\mathbb{E}f(w^t) - \mathbb{E}f\left(w^{t+1}\right)\right]}{\eta\beta Q} \tag{A70}$$

$$+ 8\eta^2L^2Q^2(1+2Q)(\tau^2+1)(1+\beta^2)(1+4\eta\beta L)\left(\sigma_a^2 + \gamma_g^2\right) \tag{A71}$$

$$+ 4\eta\beta LQ\left(\sigma_a^2 + \gamma_g^2\right). \tag{A72}$$

Now, note that for any $s \geq 0$,[6]

$$\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2 \leq \sum_{u=s-\tau}^{s}\mathbb{E}\left\|\nabla f\left(w^u\right)\right\|^2, \tag{A73}$$

[6]For $s < \tau$, the right-hand side of the inequality consists of fewer terms.

Therefore, we add up the inequality in (A68)-(A72), for $t = 0, 1, \dots T-1$, and obtain

$$\left[1 - 4\eta\beta LQ - 8\eta^2 L^2 Q^2 (1+2Q)\tau(\tau+1)^2(1+\beta^2)(1+4\eta\beta L)\right] \frac{\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(w^t)\right\|^2}{T} \tag{A74}$$

$$\leq \frac{2\left[f(w^0) - \mathbb{E}f(w^T)\right]}{\eta\beta QT} + 4\eta\beta LQ\left(\sigma_a^2 + \gamma_g^2\right) \tag{A75}$$

$$+ 8\eta^2 L^2 Q^2 (1+2Q)(\tau^2+1)(1+\beta^2)(1+4\eta\beta L)\left(\sigma_a^2 + \gamma_g^2\right). \tag{A76}$$

Thus, by setting $\beta = 1$ and $\eta = \frac{1}{Q\sqrt{LT}}$, we can simply see that

$$1 - 4\eta\beta LQ - 8\eta^2 L^2 Q^2 (1+2Q)\tau(\tau+1)^2(1+\beta^2)(1+4\eta\beta L) \geq \frac{1}{2}, \tag{A77}$$

$$\eta \leq \frac{1}{4L(Q+1)}, \tag{A78}$$

for $T \geq 160L(Q+7)(\tau+1)^3$. Therefore, we can conclude the final result in Theorem 1 under this choice of η and β. □

Appendix B. Personalized Asynchronous Federated Learning: MAML

Proof of Theorem 2. To simplify (9), we denote $\tilde{\nabla} F_i^{(b)}(w) = \nabla \tilde{F}_i^{(b)}(w, \mathcal{D}_i'', \mathcal{D}_i', \mathcal{D}_i)$. Then similar to (A5), at round t, the update rule for Option B can be written as follows:

$$w^{t+1} = w^t - \eta\beta \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right). \tag{B1}$$

According to Lemma 1,

$$F^{(b)}\left(w^{t+1}\right) \overset{(12)}{\leq} F^{(b)}(w^t) - \eta\beta \underbrace{\left\langle \nabla F^{(b)}(w^t), \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right) \right\rangle}_{=:S_{b_1}} + \frac{L_b \eta^2 \beta^2}{2} \underbrace{\left\| \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right) \right\|^2}_{=:S_{b_2}} \tag{B2}$$

Similar to the inequalities in (A7)-(A12), we first show a lower bound on term S_{b_1} in (B2). We also denote $\tilde{g}_i^t = \sum_{q=0}^{Q-1} \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right)$, $\tilde{g}^t = \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i^t$, $g_i^t = \sum_{q=0}^{Q-1} \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right)$, and $g^t = \frac{1}{n} \sum_{i=1}^{n} g_i^t$ for simplicity. Note that $\tilde{g}_i^t$ and g_i^t are the stochastic and deterministic gradients of the personalized cost functions $F_i^{(b)}$ at stale parameters. According to these definitions, we have

$$\left\| \mathbb{E}\left[\tilde{g}^t - g^t \right] \right\| \overset{(A1b)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbb{E}\left[\tilde{g}_i^t - g_i^t \right] \right\| \tag{B3}$$

$$\overset{(A1b)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \left\| \mathbb{E}\left[\tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \right] \right\| \tag{B4}$$

$$\overset{(22)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mu_b = Q\mu_b, \tag{B5}$$

where as we discussed in (22), μ_b measures the unbiasedness in the estimation of the personalized stochastic gradient.

$$\mathbb{E}\left[S_{b_1}\right] = \mathbb{E}\left[\mathbb{E}_{i_t}\left\langle \nabla F^{(b)}(w^t), \tilde{g}_{i_t}^t \right\rangle\right] \tag{B6}$$

$$= \mathbb{E}\left[\left\langle \nabla F^{(b)}(w^t), \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^t \right\rangle\right] = \mathbb{E}\left[\left\langle \nabla F^{(b)}(w^t), \tilde{g}^t \right\rangle\right] \tag{B7}$$

$$= Q\,\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 + \mathbb{E}\left\langle \nabla F^{(b)}(w^t), \mathbb{E}\left[\tilde{g}^t - g^t\right]\right\rangle \tag{B8}$$

$$+ \mathbb{E}\left\langle \nabla F^{(b)}(w^t), g^t - Q\nabla F^{(b)}(w^t)\right\rangle \tag{B9}$$

$$\overset{\text{(A1c)}}{\geq} Q\,\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 - \frac{1}{4}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 - \left\|\mathbb{E}\left[g^t - \tilde{g}^t\right]\right\|^2 \tag{B10}$$

$$- \frac{1}{4}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 - \mathbb{E}\left\|g^t - Q\nabla F^{(b)}(w^t)\right\|^2 \tag{B11}$$

$$\overset{\text{(B3)}-\text{(B5)}}{\geq} \frac{2Q-1}{2}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 - \mathbb{E}\left\|g^t - Q\nabla F^{(b)}(w^t)\right\|^2 - Q^2\mu_b^2, \tag{B12}$$

and

$$\mathbb{E}_{i_t}\left[S_{b_2}\right] = \mathbb{E}_{i_t}\left\|\sum_{q=0}^{Q-1}\tilde{\nabla}F_{i_t}^{(b)}\left(w_{i_t,q}^{\Omega(t)}\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^n\left\|\tilde{g}_i^t\right\|^2. \tag{B13}$$

Therefore, according to (B2), (B12), and (B13),

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \frac{\eta\beta(2Q-1)}{2}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 + \eta\beta Q^2\mu_b^2 \tag{B14}$$

$$+ \eta\beta\,\underbrace{\mathbb{E}\left\|g^t - Q\nabla F^{(b)}(w^t)\right\|^2}_{=:S_{b_3}} + \frac{L_b\eta^2\beta^2}{2n}\underbrace{\mathbb{E}\sum_{i=1}^n\left\|\tilde{g}_i^t\right\|^2}_{=:S_{b_4}}, \tag{B15}$$

where similar to (A16)-(A19), we can bound S_{b_3} as follows:

$$S_{b_3} \leq \frac{Q}{n}\sum_{i=1}^n\sum_{q=0}^{Q-1}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}(w^t)\right\|^2. \tag{B16}$$

Moreover, we can show an upper bound on S_{b_4} akin to (B17)-(B20):

$$S_{b_4} = \sum_{i=1}^{n} \Big\| \sum_{q=0}^{Q-1} \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \tag{B17}$$

$$\overset{\text{(A1d)}}{\leq} Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \Big\| \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \tag{B18}$$

$$= Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \Big\| \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) + \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w^t \right) + \nabla F_i^{(b)} \left(w^t \right) - \nabla F^{(b)} \left(w^t \right) + \nabla F^{(b)} \left(w^t \right) \Big\|^2 \tag{B19}$$

$$\overset{\text{(A1d)}}{\leq} 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \Bigg[\Big\| \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 + \Big\| \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w^t \right) \Big\|^2 + \Big\| \nabla F_i^{(b)} \left(w^t \right) - \nabla F^{(b)} \left(w^t \right) \Big\|^2 + \Big\| \nabla F^{(b)} \left(w^t \right) \Big\|^2 \Bigg] \Rightarrow \tag{B20}$$

$$\mathbb{E}[S_{b_4}] \overset{\text{(B20)}}{\leq} 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E}_{p_i} \left[\Big\| \tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \Big\|^2 \right] \tag{B21}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w^t \right) \Big\|^2 \tag{B22}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla F_i^{(b)} \left(w^t \right) - \nabla F^{(b)} \left(w^t \right) \Big\|^2 \tag{B23}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \Big\| \nabla F^{(b)} \left(w^t \right) \Big\|^2 \tag{B24}$$

$$\overset{(15),(16)}{\leq} 4nQ^2 \left[\sigma_b^2 + \gamma_b^2 + \mathbb{E} \Big\| \nabla F^{(b)} \left(w^t \right) \Big\|^2 \right] \tag{B25}$$

$$+ 4Q \sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E} \left\| \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} (w^t) \right\|^2. \tag{B26}$$

Therefore, due to (B14)-(B16) and (B25)-(B26), we have

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_b\beta^2Q^2\right]\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{B27}$$

$$+ \left[\frac{\eta\beta Q}{n} + \frac{2\eta^2\beta^2 Q L_b}{n}\right]\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}(w^t)\right\|^2 \tag{B28}$$

$$+ \eta\beta Q^2\mu_b^2 + 2\eta^2 L_b\beta^2Q^2\sigma_b^2 + 2\eta^2 L_b\beta^2Q^2\gamma_b^2 \tag{B29}$$

$$\overset{(12)}{\leq} \mathbb{E}F^{(b)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_b\beta^2Q^2\right]\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{B30}$$

$$+ \frac{\eta\beta Q L_b^2\left(1+2\eta\beta L_b\right)}{n}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\underbrace{\left\|w_{i,q}^{\Omega(t)} - w^t\right\|^2}_{=:S_{b_5}} \tag{B31}$$

$$+ \eta\beta Q^2\mu_b^2 + 2\eta^2 L_b\beta^2Q^2\left(\sigma_b^2 + \gamma_b^2\right). \tag{B32}$$

Now, we provide an upper bound on S_{b_5} in (B30) as follows:

$$S_{b_5} = \left\|w^t - w_{i,q}^{\Omega(t)}\right\|^2 = \left\|w^t - w^{\Omega(t)} + w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{B33}$$

$$\overset{(A1a)}{\leq} 2\underbrace{\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2}_{S_{b_6}} + 2\underbrace{\left\|w^t - w^{\Omega(t)}\right\|^2}_{S_{b_7}}, \tag{B34}$$

where the first term determines the evolution of local updates and the second term considers the effect of asynchronous updates. Therefore, using Lemma 4, we have

$$\mathbb{E}[S_{b_6}] = \mathbb{E}\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{B35}$$

$$= \mathbb{E}\left\|w_{i,0}^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{B36}$$

$$\overset{(B1)}{=} \eta^2\mathbb{E}\left\|\sum_{r=0}^{q-1}\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 \tag{B37}$$

$$\overset{(A1d)}{\leq} \eta^2 q\sum_{r=0}^{q-1}\mathbb{E}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 \tag{B38}$$

$$\overset{(A1d)}{\leq} 2\eta^2 q\sum_{r=0}^{q-1}\left[\mathbb{E}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 + \mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2\right] \tag{B39}$$

$$\overset{(23),(25)}{\leq} 2\eta^2 q\sum_{r=0}^{q-1}\left(G_b^2 + \sigma_b^2\right) = 2\eta^2q^2\left(G_b^2 + \sigma_b^2\right), \tag{B40}$$

$$\mathbb{E}[S_{b_7}] = \mathbb{E}\left\|w^t - w^{\Omega(t)}\right\|^2 \tag{B41}$$

$$= \mathbb{E}\left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1}-w^s\right)\right\|^2 \tag{B42}$$

$$\overset{\text{Alg. 1,2}}{=} \eta^2\beta^2\mathbb{E}\left\|\sum_{s=\Omega(t)}^{t-1}\sum_{q=0}^{Q-1}\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2 \tag{B43}$$

$$\overset{\text{(A1d)}}{\leq} \eta^2\beta^2 Q\left(t-\Omega(t)\right)\sum_{s=\Omega(t)}^{t-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2 \tag{B44}$$

$$\overset{(11)}{\leq} 2\eta^2\beta^2 Q\,\tau\sum_{s=t-\tau}^{t-1}\sum_{q=0}^{Q-1}\left[\mathbb{E}\left\|\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)-\nabla F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2\right.$$

$$\left.+\,\mathbb{E}\left\|\nabla F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2\right] \tag{B45}$$

$$\overset{(23),(25)}{\leq} 2\eta^2\beta^2 Q\tau^2\sum_{q=0}^{Q-1}\left(G_b^2+\sigma_b^2\right) = 2\eta^2\beta^2Q^2\tau^2\left(G_b^2+\sigma_b^2\right). \tag{B46}$$

So, according to (B27)-(B46),

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \frac{\eta\beta}{2}\left(2Q-1-4\eta\beta L_b Q^2\right)\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{B47}$$

$$+\,4\eta^3\beta Q^4 L_b^2\left(1+2\eta\beta L_b Q\right)\left(G_b^2+\sigma_b^2\right)\left(\beta^2\tau^2+1\right) \tag{B48}$$

$$+\,\eta\beta Q^2\mu_b^2 + 2\eta^2\beta^2 L_b Q^2\sigma_b^2 + 2\eta^2\beta^2 L_b Q^2\gamma_b^2, \tag{B49}$$

where by adding the terms in (B47)-(B49), for $t = 0, 1, \ldots T-1$, and rearranging them, we obtain the following inequality:

$$\frac{1-4\eta\beta L_b Q}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \leq \frac{2\left(F^{(b)}(w^0)-\mathbb{E}F^{(b)}(w^T)\right)}{\eta\beta QT} \tag{B50}$$

$$+\,8\eta^2Q^3L_b^2\left(1+2\eta\beta L_bQ\right)\left(G_b^2+\sigma_b^2\right)\left(\beta^2\tau^2+1\right) \tag{B51}$$

$$+\,4\eta\beta L_bQ\left(\sigma_b^2+\gamma_b^2\right) \tag{B52}$$

$$+\,2Q\mu_b^2. \tag{B53}$$

Finally, we can conclude the proof by fixing $\beta = 1$ and $\eta \coloneqq \frac{1}{Q\sqrt{L_bT}}$ for $T \geq 64L_b$, hence $\eta \leq \frac{1}{8\beta L_bQ}$. □

Appendix C. Personalized Asynchronous Federated Learning: ME

We start by showing (10). According to the definitions in (6) and (8), we have

$$\hat{\theta}_i(w) = \underset{\theta_i \in \mathbb{R}^d}{\arg\min} \left[f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - w\|^2 \right] \Rightarrow \nabla f_i \left(\hat{\theta}_i(w) \right) + \lambda \left[\hat{\theta}_i(w) - w \right] = 0, \quad \text{(C1)}$$

$$F_i^{(c)}(w) = f_i \left(\hat{\theta}_i(w) \right) + \frac{\lambda}{2} \left\| \hat{\theta}_i(w) - w \right\|^2, \quad \text{(C2)}$$

therefore,

$$\nabla F_i^{(c)}(w) \overset{\text{(C2)}}{=} \frac{\partial \hat{\theta}_i(w)}{\partial w} \left[\nabla f_i \left(\hat{\theta}_i(w) \right) \right] + \lambda \left[\frac{\partial \hat{\theta}_i(w)}{\partial w} - I \right] \left[\hat{\theta}_i(w) - w \right] \quad \text{(C3)}$$

$$\overset{\text{(C1)}}{=} \lambda \frac{\partial \hat{\theta}_i(w)}{\partial w} \left[w - \hat{\theta}_i(w) \right] + \lambda \left[\frac{\partial \hat{\theta}_i(w)}{\partial w} - I \right] \left[\hat{\theta}_i(w) - w \right] \quad \text{(C4)}$$

$$= \lambda \left[w - \hat{\theta}_i(w) \right]. \quad \text{(C5)}$$

Before, presenting the proof of Theorem 3, we proceed by providing the proof of Lemmas 5, 6, and 7.

Proof of Lemma 5. Let w, v be two arbitrary vectors in $\mathbb{R}^d$. Then, we have:

$$\nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \overset{\text{(C5)}}{=} \lambda \left[w - \hat{\theta}_i(w) \right] - \lambda \left[v - \hat{\theta}_i(v) \right] \quad \text{(C6)}$$

$$\overset{\text{(C1)}}{=} \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \Rightarrow \quad \text{(C7)}$$

$$\left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| = \left\| \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \quad \text{(C8)}$$

$$\overset{(12)}{\leq} L \left\| \hat{\theta}_i(w) - \hat{\theta}_i(v) \right\| \quad \text{(C9)}$$

$$\overset{\text{(C1)}}{=} L \left\| w - \frac{1}{\lambda} \nabla f_i \left(\hat{\theta}_i(w) \right) - v + \frac{1}{\lambda} \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \quad \text{(C10)}$$

$$\leq L \|w - v\| + \frac{L}{\lambda} \left\| \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \quad \text{(C11)}$$

$$= L \|w - v\| + \frac{L}{\lambda} \left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| \Rightarrow \quad \text{(C12)}$$

$$\left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| \leq \frac{\lambda L}{\lambda - L} \|w - v\|, \quad \text{(C13)}$$

which means $F_i^{(c)}$ is $\frac{\lambda L}{\lambda - L}$-smooth. Note that for $\lambda \geq \kappa L$, for some $\kappa > 1$,

$$\frac{\lambda L}{\lambda - L} \leq L_c \coloneqq \frac{\lambda}{\kappa - 1} \quad \text{(C14)}$$

This concludes the statement of Lemma 5. □

Proof of Lemma 6. According to Step 11 of Algorithm 2, let us introduce full and

stochastic auxiliary cost functions $h_i(\cdot)$ and $\tilde{h}_i(\cdot)$ as follows:

$$h_i(\theta_i, w) = f_i(\theta_i) + \frac{\lambda}{2}\left\|\theta_i - w\right\|^2, \tag{C15}$$

$$\tilde{h}_i(\theta_i, w, \mathcal{D}) = \tilde{f}_i(\theta_i, \mathcal{D}) + \frac{\lambda}{2}\left\|\theta_i - w\right\|^2, \tag{C16}$$

where due to (C15), we have

$$\nabla\tilde{h}_i(\tilde{\theta}_i(w), w, \mathcal{D}) = \nabla\tilde{f}_i(\tilde{\theta}_i(w), \mathcal{D}) + \lambda\left[\tilde{\theta}_i(w) - w\right], \tag{C17}$$

hence, we can show (26) as follows:

$$\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w, \mathcal{D}) - \nabla F_i^{(c)}(w)\right]\right\| \tag{C18}$$

$$\overset{(7),(10)}{=} \left\|\mathbb{E}_{p_i}\left[\lambda\hat{\theta}_i(w) - \lambda\tilde{\theta}_i(w)\right]\right\| \tag{C19}$$

$$\overset{(C1),(C16)}{=} \left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w)) - \nabla\tilde{f}_i(\tilde{\theta}_i(w), \mathcal{D}) + \nabla\tilde{h}_i(\tilde{\theta}_i(w), w, \mathcal{D})\right]\right\| \tag{C20}$$

$$= \left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\tilde{\theta}_i(w))\right] + \mathbb{E}_{p_i}\left[\nabla\tilde{h}_i(\tilde{\theta}_i(w), w, \mathcal{D})\right]\right\| \tag{C21}$$

$$\leq \left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\tilde{\theta}_i(w))\right]\right\| + \nu \tag{C22}$$

$$\overset{(12)}{\leq} L\left\|\mathbb{E}_{p_i}\left[\hat{\theta}_i(w) - \tilde{\theta}_i(w)\right]\right\| + \nu \tag{C23}$$

$$\overset{(C19)}{=} \frac{L}{\lambda}\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w, \mathcal{D}) - \nabla F_i^{(c)}(w)\right]\right\| + \nu \Rightarrow \tag{C24}$$

$$\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w, \mathcal{D}) - \nabla F_i^{(c)}(w)\right]\right\| \leq \frac{\lambda}{\lambda - L}\nu. \tag{C25}$$

You can find the proof of (27) in [13][Appendix A.2]. □

Proof of Lemma 7. First, note that we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\right\|^2 \tag{C26}$$

$$\overset{(C5)}{=} \frac{1}{n}\sum_{i=1}^{n}\left\|\lambda(w - \hat{\theta}_i(w)) - \frac{1}{n}\sum_{j=1}^{n}\lambda(w - \hat{\theta}_i(w))\right\|^2 \tag{C27}$$

$$\overset{(C1)}{=} \frac{1}{n^3}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}\left[\nabla f_i(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\right]\right\|^2 \tag{C28}$$

$$\overset{(A1d)}{\leq} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\right\|^2. \tag{C29}$$

So, we simplify the upper bound as follows:

$$\left\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\right\|^2 \tag{C30}$$

$$= \Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w)) + \nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w)) + \nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w)) + \nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w)) + \nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{C31}$$

$$\overset{\text{(A1a)}}{\leq} \frac{4}{3}\Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w)) + \nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w)) + \nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{C32}$$

$$+ 4\Big\|\nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w)) + \nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \tag{C33}$$

$$\overset{\text{(A1d)}}{\leq} 4\Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w))\Big\|^2 \tag{C34}$$

$$+ 4\Big\|\nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w))\Big\|^2 \tag{C35}$$

$$+ 4\Big\|\nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{C36}$$

$$+ 8\Big\|\nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \tag{C37}$$

$$+ 8\Big\|\nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w))\Big\|^2 \tag{C38}$$

Note that we can bound (C37) and (C38) according to Lemma 16:

$$\frac{1}{n}\sum_{i=1}^{n}\Big\|\nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \overset{(16)}{\leq} \gamma_g^2, \tag{C39}$$

and also given the fact that function $f(\cdot)$ as well as each function $f_i(\cdot)$ are L-smooth, we can bound (C34), (C35), and (C36) as follows:

$$\left\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w))\right\|^2 \tag{C40}$$

$$\leq L^2\Big\|\hat{\theta}_i(w) - \hat{\theta}_j(w)\Big\|^2 \tag{C41}$$

$$= \frac{L^2}{\lambda^2}\Big\|\lambda\left[\hat{\theta}_i(w) - w\right] - \lambda\left[\hat{\theta}_j(w) - w\right]\Big\|^2 \tag{C42}$$

$$\overset{\text{(C5)}}{=} \frac{L^2}{\lambda^2}\Big\|\nabla F_i^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2 \tag{C43}$$

$$= \frac{L^2}{\lambda^2}\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w) + \nabla F^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2 \tag{C44}$$

$$\overset{\text{(A1d)}}{\leq} \frac{2L^2}{\lambda^2}\left[\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\Big\|^2 + \Big\|\nabla F^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2\right]. \tag{C45}$$

Therefore, according to (C46)-(C45), we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w)-\nabla F^{(c)}(w)\right\|^2 \leq 16\gamma_g^2+\frac{48L^2}{n\lambda^2}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w)-\nabla \quad F^{(c)}(w)\right\|^2 \Rightarrow \tag{C46}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w)-\nabla F^{(c)}(w)\right\|^2 \leq \frac{16\lambda^2\gamma_g^2}{\lambda^2-48L^2}, \tag{C47}$$

which concludes the proof. □

Now, we are ready to state the proof of Theorem 3.

Proof of Theorem 3. We write $\tilde{\nabla}F_i^{(c)}(w)=\nabla\tilde{F}_i^{(c)}(w,\mathcal{D}_i)$ to simplify (10). Then, the update rule for Algorithms 1 & 2 under Option C can be written as follows:

$$w^{t+1}=w^t-\eta\beta\sum_{q=0}^{Q-1}\tilde{\nabla}F_{i_t}^{(c)}\left(w_{i_t,q}^{\Omega(t)}\right), \tag{C48}$$

where similar to (B2)-(B32), we can show that:

$$\mathbb{E}F^{(c)}\left(w^{t+1}\right)\leq \mathbb{E}F^{(c)}(w^t)-\left[\frac{\eta\beta(2Q-1)}{2}-2\eta^2L_c\beta^2Q^2\right]\mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2 \tag{C49}$$

$$+\frac{\eta\beta QL_c^2\,(1+2\eta\beta L_c)}{n}\underbrace{\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|w_{i,q}^{\Omega(t)}-w^t\right\|^2}_{=:S_{c_1}} \tag{C50}$$

$$+\eta\beta Q^2\mu_c^2+2\eta^2L_c\beta^2Q^2\left(\sigma_c^2+\gamma_c^2\right), \tag{C51}$$

with $L_c,\mu_c,\sigma_c,\gamma_c$ as defined in Lemmas 5, 6, and 7. Thus, to show the convergence rate of our method for the cost function in (6), it would only be sufficient to provide an upper bound on S_{c_1}. First, note that similar to (A36)-(A40), we have

$$\left\|w_{i,q}^{\Omega(t)}-w^t\right\|^2\leq\tau\left(1+\frac{1}{\beta^2}\right)\left[\sum_{s=t-\tau}^{t-1}\underbrace{\left\|w^{s+1}-w^s\right\|^2}_{=:S_{c_3}}\right]+\left(1+\beta^2\right)\underbrace{\left\|w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^2}_{=:S_{c_2}}. \tag{C52}$$

Now, if we introduce stepsize η such that $\eta\leq\frac{1}{4L_c(Q+1)}$, similar to (A41)-(A46) and (A54)-(A58), the following two inequalities holds for S_{c_2} and S_{c_3}:

$$\mathbb{E}[S_{c_2}]=\mathbb{E}\left\|w_{i,q}^{\Omega(t)}-w^{\Omega(t)}\right\|^2 \tag{C53}$$

$$\leq 8Q(1+2Q)\eta^2\left[\sigma_c^2+\mathbb{E}\left\|\nabla F_i^{(c)}\left(w^{\Omega(t)}\right)-\nabla F^{(c)}\left(w^{\Omega(t)}\right)\right\|^2+\mathbb{E}\left\|\nabla F^{(c)}\left(w^{\Omega(t)}\right)\right\|^2\right],$$

$$\mathbb{E}\left[S_{c_3}\right] = \mathbb{E}\left\|w^{s+1} - w^s\right\|^2 \leq 8Q(1{+}2Q)\eta^2\beta^2\left[\sigma_c^2 + \gamma_c^2 + \mathbb{E}\left\|\nabla F^{(c)}\left(w^{\Omega(s)}\right)\right\|^2\right], \quad \text{(C54)}$$

where by denoting $\phi = 8\eta^2Q^2(1{+}2Q)(1{+}\beta^2)$, we have

$$\frac{1}{n\phi}\mathbb{E}[S_{c_1}] \leq (\tau^2{+}1)\left[\sigma_c^2 + \gamma_c^2\right] + \tau\sum_{s=t-\tau}^{t}\sum_{u=s-\tau}^{s}\mathbb{E}\left\|\nabla F^{(c)}\left(w^u\right)\right\|^2. \quad \text{(C55)}$$

Then, according to (C49)-(C55), we obtain

$$\begin{aligned}
\mathbb{E}F^{(c)}(w^{t+1}) \leq{}& \mathbb{E}F^{(c)}(w^t) - \left[\frac{\eta\beta(2Q{-}1)}{2} - 2\eta^2L_c\beta^2Q^2\right]\mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2 && \text{(C56)}\\
&+ 8\eta^3\beta Q^3L_c^2(1{+}2Q)(1{+}\beta^2)\left(1{+}2\eta\beta L_c\right)\tau\left[\sum_{s=t-\tau}^{t}\sum_{u=s-\tau}^{s}\mathbb{E}\left\|\nabla F^{(c)}(w^u)\right\|^2\right] && \text{(C57)}\\
&+ 8\eta^3\beta Q^3L_c^2(1{+}2Q)(\tau^2{+}1)(1{+}\beta^2)\left(1{+}2\eta\beta L_c\right)\left(\sigma_c^2 + \gamma_c^2\right) && \text{(C58)}\\
&+ 2\eta^2L_c\beta^2Q^2\left(\sigma_c^2 + \gamma_c^2\right) && \text{(C59)}\\
&+ \eta\beta Q^2\mu_c^2, && \text{(C60)}
\end{aligned}$$

where by averaging the terms in (C56)-(C60), for $t = 0, 1, \ldots T{-}1$, and rearranging them (similar to (A74)-(A76), we can conclude the following inequality:

$$\begin{aligned}
&\frac{1 - 4\eta\beta QL_c - 16\eta^2Q^2L_c^2(1{+}2Q)\tau(\tau{+}1)^2(1{+}\beta^2)\left(1{+}2\eta\beta L_c\right)}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2\\
&\leq \frac{2\left(F^{(c)}(w^0) - \mathbb{E}F^{(c)}(w^T)\right)}{\eta\beta QT} + 2Q\mu_c^2 && \text{(C61)}\\
&+ 16\eta^2Q^2L_c^2(1{+}2Q)(\tau^2{+}1)(1{+}\beta^2)\left(1{+}2\eta\beta L_c\right)\left(\sigma_c^2 + \gamma_c^2\right) && \text{(C62)}\\
&+ 4\eta\beta QL_c\left(\sigma_c^2 + \gamma_c^2\right) && \text{(C63)}
\end{aligned}$$

Finally, by fixing $\eta = \frac{1}{Q\sqrt{L_cT}}$, for $T \geq 288L_c(Q{+}7)(\tau{+}1)^2$, we obtain the sublinear convergence rate in Theorem 3. □

Appendix D. Experiments Setting

For all algorithms, we consider $Q = 10$ local updates, and select the best $\lambda \in \{20, 25, 30\}$ for `ME` and $\alpha \in \{0.002, 0.005, 0.01\}$ for `MAML`. Moreover, we pick $\beta \in \{0.8, 1.0, 1.2\}$ and fix $\eta = 0.01$. For all experiments, we consider the exact same communication setup and repeat each experiment 2 to 3 times and plot the test accuracy curve over time until one of the algorithms converges. We consider $n = 30$ agents for all experiments. Moreover, for both datasets we consider ℓ-layer CNN networks [39] followed by ℓ-fully connected layers with pooling and dropout as well as cross-entropy loss, where $\ell = 2$ for MNIST and $\ell = 3$ for CIFAR-10. Also, for MNIST we consider $c = 5$ class of samples for each client while for CIFAR we create heterogeneity by considering $c = 3$ per client.

It is worth mentioning that in the implementation of algorithms with `MAML`, we approximated the Hessian-vector products via the following first-order formulation: for some small $\delta > 0$,

$$\nabla^2 f_i(w)u \approx \frac{\nabla f_i(w + \delta u) - \nabla f_i(w - \delta u)}{\delta}. \tag{D1}$$

Moreover, in the bi-level optimization problem for the `ME` formulation, we applied a constant $K = 10$ steps of `SGD` to obtain $\tilde{\theta}_i(w)$.