PersA-FL: Personalized Asynchronous Federated Learning

Mohammad Taha Toghani
Rice University
Houston, TX, USA
mttoghani@rice.edu

Soomin Lee
Yahoo! Research
Sunnyvale, CA, USA

César A. Uribe
Rice University
Houston, TX, USA
cauribe@rice.edu

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. Our main technical contribution is a unified proof for asynchronous federated learning with bounded staleness that we apply to MAML and ME frameworks. For the smooth and non-convex functions class, we show the convergence of our method to a first-order stationary point.

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 [19, 15]. The common underlying assumption that determines the superiority of vanilla 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. 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. Fallah et al. [8] 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. [6] present pFedMe algorithm for PFL via adopting a different formulation for personalization, namely Moreau Envelopes (ME). Several recent works have approached PFL mainly through optimization-based [13, 23, 37, 27, 5, 12], or structure-based [4, 32] techniques.

In cross-device FL, devices are naturally prone to update and communicate models under less restrictive rules, whereas clients may apply updates in an asynchronous fashion, i.e., staleness. Hogwild! [26] 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 [22]. 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 [1, 18]. More closely, FL under stale updates has been thoroughly studied in [35, 25, 2].

In this work, we study the PFL problem under asynchronous communications to improve training concurrency, performance, and efficiency. Through the integration of two personalization formulations, MAML & ME, we propose PersA-FL, an algorithm that allows PFL under asynchronous communications with the server. We present the client algorithm under three different options for the

6th Workshop on Meta-Learning at NeurIPS 2022, New Orleans.

local updates, each addressing a separate formulation, (A) vanilla AFL, (B) PersA-FL: MAML, and (C) PersA-FL: ME. We show the convergence rate of PersAFL based on the maximum delay and personalization budget.

2 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 obtain a model $w \in \mathbb{R}^d$ that

$$\min_{w \in \mathbb{R}^d} f(w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f_i(w), \tag{1}$$

with $f_i(w) \coloneqq \mathbb{E}_{\Xi_i \sim p_i}[\ell_i(w, \Xi_i)]$, 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 $i \in [n]$. Let $\mathcal{D}_i$ be a data batch with samples 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. 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 an alternative problem formulation for (1).

2.1 Personalized Federated Learning

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 vanilla 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. [8] 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, [8] define the MAML loss function for PFL as follows:

$$\min_{w \in \mathbb{R}^d} F^{(b)}(w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} F_i^{(b)}(w), \tag{3}$$

with $F_i^{(b)}(w) \coloneqq f_i(w - \alpha \nabla f_i(w))$ where $\alpha \geq 0$ is the MAML personalization stepsize. Solving (3) 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 (3), stepsize α determines the personalization budget, where $\alpha = 0$ implies vanilla FL in Problem (1). See [14, 31, 9] 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 (3), which can be written as follows:

$$\nabla F_i^{(b)}(w) = \left[I - \alpha \nabla^2 f_i(w)\right] \nabla f_i\left(w - \alpha \nabla f_i(w)\right), \tag{4}$$

where in Per-FedAvg, the authors propose to compute a biased estimation of (4) using stochastic gradients/Hessian. We will elaborate on the stochastic approximation in Section 3.

3 Algorithm & Convergence Result

In this section, we propose Algorithms 1 & 3 to solve Problem 3. We move the result ME to Appendix A. We present our method through two different perspectives, (i) server and (ii) client.

$\diamond$ **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), (3), or (8). 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

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

1: **input:** number of local steps Q, local stepsize η, `MAML` stepsize α, Moreau Envelope (ME) regularization parameter λ, minimum batch size b, estimation error ν.
2: **repeat**
3: read w from the server
4: $w_{i,0} \leftarrow w$
5: **for** $q = 0$ to $Q-1$ **do**
6: sample a data batch $\mathcal{D}_{i,q}$ from distribution p_i
▷ **Option A** (AFL)
▷ **Option B** (`PersA-FL: MAML`)
7: sample two data batches $\mathcal{D}'_{i,q}, \mathcal{D}''_{i,q}$ from distribution p_i
8: $w_{i,q+1} \leftarrow w_{i,q} - \eta \left[I - \alpha \nabla^2 \tilde{f}_i(w_{i,q}, \mathcal{D}''_{i,q})\right] \nabla \tilde{f}_i \left(w_{i,q} - \alpha \nabla \tilde{f}_i(w_{i,q}, \mathcal{D}'_{i,q}), \mathcal{D}_{i,q}\right)$
▷ **Option C** (`PersA-FL: ME`)
9: **end for**
10: $\Delta_i \leftarrow w_{i,0} - w_{i,Q}$
11: client i broadcasts Δ_i to the server
12: **until** not interrupted by the server

Note that we drop the time index from the iterates of the client algorithm for clarity of exposition.

$\diamond$ **Client Algorithm**: 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), (3), or (8), as in Steps 5-9, and (iii) uploading the sum of updates on the server as in Step 11. 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. By performing this option, we aim to minimize the `MAML` cost function in (3). As we saw in Section 2, the full gradient can be computed according to (4). Following [8], we sample three data batches to compute a biased estimation of (4) as follows:

$$\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}''_i, \mathcal{D}'_i, \mathcal{D}_i) = \left[I - \alpha \nabla^2 \tilde{f}_i(w, \mathcal{D}''_i)\right] \nabla \tilde{f}_i \left(w - \alpha \nabla \tilde{f}_i(w, \mathcal{D}'_i), \mathcal{D}_i\right). \tag{5}$$

We will discuss the variance and bias of this estimator in Subsection B.2. Next, we present the convergence result of our method for the three formulations.

3.1 Convergence Results

Now, we show the convergence of our method for MAML setup. 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 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.,*

$$\sup_{t \geq 0} |t - \Omega(t)| \leq \tau, \tag{6}$$

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

Assumption 2 (Local Cost Properties)**.** *For all clients $i \in [n]$, function $f_i : \mathbb{R}^d \to \mathbb{R}$ is bounded below, twice differentiable, and the following properties hold for all $w, u \in \mathbb{R}^d$,*

$$\begin{cases} \|\nabla f_i(w) - \nabla f_i(u)\| \leq L\|w - u\|, \\ \mathbb{E}_{\xi_i \sim p_i} \|\nabla \ell_i(w, \xi_i) - \nabla f_i(w)\|^2 \leq \sigma_g^2, \\ \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \gamma_g^2, \\ \|\nabla f_i(w)\| \leq G, \end{cases} \quad \begin{cases} \left\|\nabla^2 f_i(w) - \nabla^2 f_i(u)\right\| \leq \rho\|w - u\|, \\ \mathbb{E}_{\xi_i \sim p_i} \left\|\nabla^2 \ell_i(w, \xi_i) - \nabla^2 f_i(w)\right\|^2 \leq \sigma_h^2, \\ \frac{1}{n} \sum_{i=1}^{n} \left\|\nabla^2 f_i(w) - \nabla^2 f(w)\right\|^2 \leq \gamma_h^2. \end{cases} \tag{7}$$

Assumption 2 summarizes all first-order and second order assumptions in [8]. We elaborate on all in Appendix B.

Theorem 1 (PersA-FL: MAML)**.** *Let Assumptions 1 and 2 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 & 3 under Option B on Problem* (3)*: for any timestep $T \geq 64 L_b$ at the server*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\|\nabla F^{(b)}(w^t)\right\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{\alpha^2 \sigma_g^2}{b}\right).$$

The proof of this theorem can be found in Appendix D. 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 updates. For instance, under a fixed personalization budget $\alpha \geq 0$, our method requires $T = \mathcal{O}(\varepsilon^{-2})$ and $b = \mathcal{O}(\varepsilon^{-1})$ to reach an ε-approximate first-order stationary solution. The last expression can also be controlled through a combined stepsize α and batch size b. It is consistent with intuition, meaning more samples are needed to obtain a higher degree of personalization. We present a detailed analysis of our method in Appendix B.

4 Conclusion

This work studied the personalized federated learning problem for the heterogeneous data setting under asynchronous communications with the server. We considered the MAML and 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. The extensions of our method to the buffered aggregation setups remain for future studies.

Acknowledgments and Disclosure of Funding

This work was partly done while Mohamamd Taha Toghani interning at Yahoo! Research. Part of this material is based upon work supported by the National Science Foundation under Grants #2211815 and #2213568.

References

[1] 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.

[2] 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.

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

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Abdurakhmon Sadiev, Ekaterina Borodich, Aleksandr Beznosikov, Darina Dvinskikh, Saveliy Chezhegov, Rachael Tappenden, Martin Takáč, and Alexander Gasnikov. Decentralized personalized federated learning: Lower bounds and optimal algorithm for all personalization modes. *EURO Journal on Computational Optimization*, page 100041, 2022.

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

[29] 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.

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

[31] 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. doi: 10.1109/LCSYS.2022.3189317.

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

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

A Problem Setup, Algorithm, and Comparisons: Extension

As an alternative option to MAML formulation in (3), Dinh et al. [6] suggest solving the following optimization problem:

$$\begin{aligned} &\min_{w\in\mathbb{R}^d} F^{(c)}(w) \coloneqq \frac{1}{n}\sum_{i=1}^{n} F_i^{(c)}(w), \\ &\text{with} \quad F_i^{(c)}(w) \coloneqq \min_{\theta_i\in\mathbb{R}^d} \left[f_i(\theta_i) + \frac{\lambda}{2}\|\theta_i - w\|^2 \right], \end{aligned} \tag{8}$$

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 (8) converges to vanilla FL in (1) with no personalization which is similar to the case in (3) 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 E to see the proof):

$$\nabla F_i^{(c)}(w) = \lambda \left(w - \hat{\theta}_i(w) \right), \tag{9}$$

$$\text{with} \quad \hat{\theta}_i(w) \coloneqq \arg\min_{\theta_i\in\mathbb{R}^d} \left[f_i(\theta_i) + \frac{\lambda}{2}\|\theta_i - w\|^2 \right], \tag{10}$$

where for large λ, $\hat{\theta}_i(w)$ is the exact solution to an optimization problem. Therefore, solving (8) through a similar approach to FedAvg or Per-FedAvg, itself requires minimizing Problem (10) which is potentially intractable. Dinh et al. [6] propose a bi-level optimization algorithm called pFedMe, to minimize the optimization problem in (8) by alternating minimization over θ_i and w. The main idea behind pFedMe is to integrating the computation of an inexact solution to (10) inside an FL-type method. We will explain and use this inexact approximation in the presentation of our method (Option C) in Section 3.

A.1 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 **??**, 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.

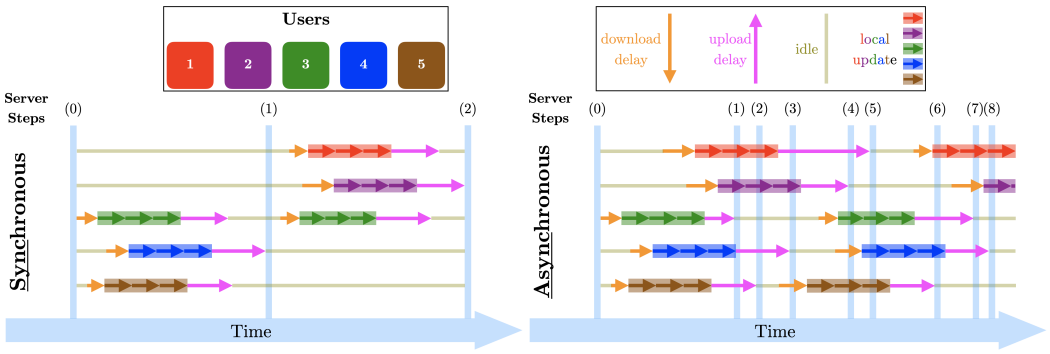


Figure 1: Communication and update schedule for synchronous and asynchronous aggregation: The demonstrated setup in this example contains $n = 5$ clients with $Q = 3$ local updates.

A.2 Algorithm: Extension

Here, we present the client algorithm consisting all options.

$\diamond$ **Client Algorithm**: Let us explain the operations of i-th client using the pseudo code in Algorithm 3. 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), (3), or (8), 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 vanilla loss as in (2).
- **Option B:** By performing this option, we aim to minimize the MAML cost function in (3). As we saw in Section 2, the full gradient can be computed according to (4). Following [8], we sample three data batches to compute a biased estimation of (4) as follows:

$$\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}_i'', \mathcal{D}_i', \mathcal{D}_i) = \left[I - \alpha \nabla^2 \tilde{f}_i(w, \mathcal{D}_i'')\right] \nabla \tilde{f}_i \left(w - \alpha \nabla \tilde{f}_i(w, \mathcal{D}_i'), \mathcal{D}_i\right). \tag{11}$$

  We will discuss the variance and bias of this estimator in Subsection B.2
- **Option C:** Finally, we invoke this option to minimize the ME personalized loss in (8). As we mentioned earlier, the full gradient of this cost is (9), where for a fixed w, we may obtain $\hat{\theta}_i(w)$ by minimizing (10). Instead, following [6], we define the 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 (9) with the following estimator:

$$\nabla \tilde{F}_i^{(c)}(w, \mathcal{D}_i) = \lambda \left(w - \tilde{\theta}_i(w)\right). \tag{12}$$

  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 [6]. Then according to the property of [3, 6], 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 (12) in Subsection B.3.

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

A.3 Comparison

Table 1 illustrates the properties of our proposed method and provides a comparison between our algorithm and underlying analysis with some related seminal works. As shown in this table, building

Algorithm 3 [Personalized] Asynchronous Federated Learning (**Client** i)

input: number of local steps Q, local stepsize η, `MAML` stepsize α, Moreau Envelope (ME) regularization parameter λ, minimum batch size b, estimation error ν.
repeat
  read w from the server ▷ download phase
  $w_{i,0} \leftarrow w$
  for $q = 0$ to $Q-1$ **do** ▷ local updates
    sample a data batch $\mathcal{D}_{i,q}$ from distribution p_i ▽ 3 options:
    ▷ **Option A** (AFL)
    $w_{i,q+1} \leftarrow w_{i,q} - \eta \nabla \tilde{f}_i(w_{i,q}, \mathcal{D}_{i,q})$
    ▷ **Option B** (`PersA-FL: MAML`)
    sample two data batches $\mathcal{D}'_{i,q}, \mathcal{D}''_{i,q}$ from distribution p_i
    $w_{i,q+1} \leftarrow w_{i,q} - \eta \left[I - \alpha \nabla^2 \tilde{f}_i(w_{i,q}, \mathcal{D}''_{i,q})\right] \nabla \tilde{f}_i \left(w_{i,q} - \alpha \nabla \tilde{f}_i(w_{i,q}, \mathcal{D}'_{i,q}), \mathcal{D}_{i,q}\right)$
    ▷ **Option C** (`PersA-FL: ME`)
    $\tilde{h}_i(\theta_i, w_{i,q}, \mathcal{D}_{i,q}) := \tilde{f}_i(\theta_i, \mathcal{D}_{i,q}) + \frac{\lambda}{2} \|\theta_i - w_{i,q}\|^2$
    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 $\tilde{\theta}_i(w_{i,q})$:
    $$\left\| \nabla \tilde{h}_i \left(\tilde{\theta}_i(w_{i,q}), w_{i,q}, \mathcal{D}_{i,q} \right) \right\| \leq \nu$$
    $w_{i,q+1} \leftarrow w_{i,q} - \eta \lambda (w_{i,q} - \tilde{\theta}_i(w_{i,q}))$
  end for
  $\Delta_i \leftarrow w_{i,0} - w_{i,Q}$
  client i broadcasts Δ_i to the server ▷ upload phase
until not interrupted by the server

upon the results in [8, 6], 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 **??**.

Table 1: Comparison of the characteristics considered in our work with previous methods for federated learning with guaranteed convergence 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. [24]	✗	✗	-	No Analysis
	Yu et al. [36]	✗	✗	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
	Wang et al. [33]	✗	✗	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
`FedAsync`	Xie et al. [35]	✗	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
`FedBuff`	Nguyen et al. [25]	✗	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
`Per-FedAvg`	Fallah et al. [8]	✓	✗	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\alpha^2}{b}\right)$
`pFedMe`	Dinh et al. [6]	✓	✗	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\lambda^2\left(\frac{1}{b}+\nu^2\right)}{(\lambda-L)^2}\right)$
This Work	`AFL`	✗	✓	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)$
	`PersA-FL: MAML`	✓	✓	✗	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)+\mathcal{O}\left(\frac{\alpha^2}{b}\right)$
	`PersA-FL: ME`	✓	✓	✓	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left(\frac{\tau^2}{T}\right)+\mathcal{O}\left(\frac{\lambda^2}{(\lambda-L)^2}\nu^2\right)$

B 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 3. As mentioned earlier, we require some additional assumptions to show the convergence of `MAML`, which we will introduce in Subsection B.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 3 (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.,*

$$\sup_{t \geq 0} |t - \Omega(t)| \leq \tau, \tag{13}$$

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

The above assumption is standard in the analysis of asynchronous methods, specifically in heterogeneous settings [25, 35, 2, 18, 29, 1]. Assumption 3 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 4 (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 $w, u \in \mathbb{R}^d$,*

$$\|\nabla f_i(w) - \nabla f_i(u)\| \leq L\|w - u\| \tag{14}$$

$$f_i^\star := \min_{w \in \mathbb{R}^d} f_i(w) > -\infty. \tag{15}$$

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 5 (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 $w \in \mathbb{R}^d$*

$$\mathbb{E}_{\xi_i \sim p_i} \left\|\nabla \ell_i(w, \xi_i) - \nabla f_i(w)\right\|^2 \leq \sigma_g^2. \tag{16}$$

Assumption 5 is standard in the analysis of `SGD`-based methods and has been used in many relevant works [30, 25, 16, 33, 17, 18, 31]. 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:

$$\mathbb{E}_{p_i} \left\|\nabla \tilde{f}_i(w, \mathcal{D}_i) - \nabla f_i(w)\right\|^2 \leq \frac{\sigma_g^2}{|\mathcal{D}_i|} \leq \sigma_a^2 := \frac{\sigma_g^2}{b} \tag{17}$$

Next, we present the bounded heterogeneity assumption.

Assumption 6 (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:*

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \gamma_g^2. \tag{18}$$

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.

[8] 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 & 3) under Option A and Option C. Therefore, we present the convergence analyses starting from our results on AFL.

B.1 Asynchronous Federated Learning (Option A)

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

Theorem 2 (AFL). *Let Assumptions 3-6 hold, $\beta = 1$, and $\eta = \frac{1}{Q\sqrt{LT}}$. Then, the following property holds for the joint iterates of Algorithms 1 & 3 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 2 is provided in Appendix C. 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 AFL. Our analysis removes the unnecessary boundedness assumption on the gradient norm.

Remark 1. *Selecting $\beta = 1$ in Theorem 2, 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 [25].*

Next, we present the convergence of PersA-FL: MAML along with some technical lemmas borrowed from [8].

B.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 (3). Accordingly, we consider similar assumptions for the second-order derivatives as Assumptions 4-6.

Assumption 7 (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 $w, u \in \mathbb{R}^d$,*

$$\left\|\nabla^2 f_i(w) - \nabla^2 f_i(u)\right\| \leq \rho\|w-u\|, \tag{19}$$

$$\mathbb{E}_{\xi_i\sim p_i}\left\|\nabla^2\ell_i(w,\xi_i) - \nabla^2 f_i(w)\right\|^2 \leq \sigma_h^2, \tag{20}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla^2 f_i(w) - \nabla^2 f(w)\right\|^2 \leq \gamma_h^2. \tag{21}$$

Assumption 7 is conventional in the analysis of methods with access to second-order information [7, 8, 28, 31]. Finally, we adopt another assumption from [11, 8, 10] on the gradient norm to simplify the analysis for the MAML cost.

Assumption 8 (Bounded-Gradient). *There exists a constant G such that for all clients $i \in [n]$, and any parameter $w \in \mathbb{R}^d$,*

$$\|\nabla f_i(w)\| \leq G. \tag{22}$$

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 7-8 only

in the analysis of PersA-FL (Algorithms 1 & 3) under Option B. Under Assumptions 4 and 8, the properties in (21) and (18) can be simply derived with $\gamma_h = 2L$ and $\gamma_g = 2G$ [8].

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

Lemma 1 ([8], Lemma 4.2 - Smoothness: MAML)**.** *Let Assumptions 4 and 8 hold. Then, $F_i^{(b)}$ in (3) is L_b-smooth, i.e., for all clients $i \in [n]$, and any parameters $w, u \in \mathbb{R}^d$,*

$$\left\|\nabla F_i^{(b)}(w) - \nabla F_i^{(b)}(u)\right\| \leq L_b \|w - u\|, \tag{23}$$

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

Lemma 1 indicates that the personalized cost in (3) 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 [31][Lemma 3].

Lemma 2 ([8], Lemma 4.3 - Bounded Variance: MAML)**.** *Let Assumptions 4, 5, 7, and 8 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 $\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D})$:*

$$\left\|\mathbb{E}_{p_i}\left[\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D}) - \nabla F_i^{(b)}(w)\right]\right\| \leq \mu_b \coloneqq \frac{\alpha L(1+\alpha L)\sigma_g}{\sqrt{b}}, \tag{24}$$

$$\mathbb{E}_{p_i}\left\|\nabla \tilde{F}_i^{(b)}(w, \mathcal{D}'', \mathcal{D}', \mathcal{D}) - \nabla F_i^{(b)}(w)\right\|^2 \leq \sigma_b^2, \tag{25}$$

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 (11) is a biased estimation of the full gradient 4. The biasness is controlled by two factors, personalization stepsize α, and batch size b.[1] 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 5 on the vanilla cost; we have a tight variance based on α and b.

Lemma 3 ([8], 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:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(b)}(w) - \nabla F^{(b)}(w)\right\|^2 \leq \gamma_b^2 \coloneqq 12(1+\alpha L)^2\left[1+\alpha^2 L^2\right]\gamma_g^2 + 12\alpha^2 G^2\gamma_h^2. \tag{26}$$

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 $w \in \mathbb{R}^d$,*

$$\left\|\nabla F_i^{(b)}(w)\right\| \leq G_b \coloneqq (1+\alpha L)G. \tag{27}$$

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-(27), we are now ready to present the convergence result for PersA-FL: MAML.

[1] It should be noted that the batch size in the upper bound of (24) refers to the size of $|\mathcal{D}'|$. Recall that we use this batch to approximate the inner gradient in 12.

Theorem 3 (PersA-FL: MAML). *Let Assumptions 3-8 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 & 3 under Option B on Problem (3): 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 D. Theorem 3 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 AFL 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 updates. For instance, under a fixed personalization budget $\alpha \geq 0$, our method requires $T = \mathcal{O}(\varepsilon^{-2})$ and $b = \mathcal{O}(\varepsilon^{-1})$ to reach an ε-approximate first-order stationary solution. The last expression can also be controlled through a combined stepsize α and batch size b. It is consistent with intuition, meaning more samples are needed to obtain a higher degree of personalization.

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

B.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 (8) 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 E.

First, we present the smoothness property of ME loss.

Lemma 5 (Smoothness: ME). *Let Assumption 4 holds and $\lambda \geq \kappa L$ for some $\kappa > 1$. Then, $F_i^{(c)}$ in (8) 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 E, 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 vanilla FL. The smoothness property in Lemma 5 is tight because, $L_c \to L$ if $\lambda \to \infty$.

Corollary 1 ([6], Proposition 1). *If $\lambda \geq 2L$, then Lemma 5 implies that $F_i^{(c)}$ in (8) is λ-smooth.*

Lemma 6 (Bounded Variance: ME). *Let Assumptions 4 and 5 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 $w \in \mathbb{R}^d$,*

$$\left\|\mathbb{E}_{p_i}\left[\nabla \tilde{F}_i^{(c)}(w,\mathcal{D}) - \nabla F_i^{(c)}(w)\right]\right\| \leq \mu_c \coloneqq \frac{\lambda}{\lambda - L}\nu, \tag{28}$$

$$\mathbb{E}_{p_i}\left\|\nabla \tilde{F}_i^{(c)}(w,\mathcal{D}) - \nabla F_i^{(c)}(w)\right\|^2 \leq \sigma_c^2 \coloneqq \frac{2\lambda^2}{(\lambda-L)^2}\left[\frac{\sigma_g^2}{b} + \nu^2\right]. \tag{29}$$

This lemma is analogous to Lemma 2 in Subsection B.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 (17).

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:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w)-\nabla F^{(c)}(w)\right\|^2 \leq \gamma_c^2 \coloneqq \frac{16\lambda^2}{\lambda^2-48L^2}\gamma_g^2. \tag{30}$$

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 4, we consider bounded population diversity as in Assumption 6, average bounded diversity. [6][Assumption 3] and [34][6.1.1 Assumptions and Preliminaries, (vii)] consider a slightly stronger version of this assumption, namely uniformly "bounded heterogeneity" which is defined as follows:*

$$\max_{i\in[n]}\sup_{w\in\mathbb{R}^d}\|\nabla f_i(w)-\nabla f(w)\|^2 \leq \gamma_g^2. \tag{31}$$

Under the modified assumption in (31)*, 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 3-6

Theorem 4 (PersA-FL: ME)**.** *Let Assumptions 3-6 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 & 3 under Option C on Problem* (8)*: for any timestep* $T \geq 288L_c(Q+7)(\tau+1)^2$ *at the server*

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

This theorem 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 vanilla AFL with no personalization. Let us compare this convergence rate with the rate of pFedMe [6] 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 $\lambda \to \infty$, the last term does not vanish unless with large data batches, i.e., $b = \mathcal{O}(\varepsilon)$. Therefore, from the personalization perspective, our analysis provides a tighter bound compared to pFedMe.

C Asynchronous Federated Learning

Proof of Theorem 2. 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 $i, j \in [m]$:

$$\|w_i + w_j\|^2 \leq (1+\alpha)\|w_i\|^2 + (1+\alpha^{-1})\|w_j\|^2, \tag{32a}$$

$$\|w_i + w_j\| \leq \|w_i\| + \|w_j\|, \tag{32b}$$

$$2\langle w_i, w_j\rangle \leq \alpha\|w_i\|^2 + \alpha^{-1}\|w_j\|^2, \tag{32c}$$

$$\left\|\sum_{i=1}^m w_i\right\|^2 \leq m\left(\sum_{i=1}^m \|w_i\|^2\right). \tag{32d}$$

Now, let us rewrite the update rule of the joint iterates in Algorithms 1 & 3 Option A at time t as follows:

- Client update:

$$w_{i,0}^t = w^t, \tag{33}$$

$$w_{i,q+1}^t = w_{i,q}^t - \eta\nabla\tilde{f}_i(w_{i,q}^t, \mathcal{D}_{i,q}^t), \tag{34}$$

- Server update:

$$w^{t+1} = w^t - \beta\Delta_{i_t} = w^t - \eta\beta\sum_{q=0}^{Q-1}\nabla\tilde{f}_{i_t}\left(w_{i_t,q}^{\Omega(t)}, \mathcal{D}_{i,q}^{\Omega(t)}\right). \tag{35}$$

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:

$$w^{t+1} = w^t - \eta\beta\sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right). \tag{36}$$

Moreover, Due to Assumption 4, we can infer that f is L-smooth, thus

$$f\left(w^{t+1}\right) \overset{(14)}{\leq} f(w^t) - \eta\beta\underbrace{\left\langle\nabla f(w^t), \sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\rangle}_{=:S_{a_1}} + \frac{L\eta^2\beta^2}{2}\underbrace{\left\|\sum_{q=0}^{Q-1}\tilde{\nabla} f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\|^2}_{=:S_{a_2}} \tag{37}$$

First, we provide a lower bound on term S_{a_1} in (37). 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,

$$\mathbb{E}[S_{a_1}] = \mathbb{E}\left[\mathbb{E}_{i_t}\left\langle\nabla f(w^t), \tilde{g}_{i_t}^t\right\rangle\right] \tag{38}$$

$$= \mathbb{E}\left[\left\langle\nabla f(w^t), \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^t\right\rangle\right] \tag{39}$$

$$= \mathbb{E}\left\langle\nabla f(w^t), \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p_i}\left[\tilde{g}_i^t\right]\right\rangle = \mathbb{E}\left\langle\nabla f(w^t), \frac{1}{n}\sum_{i=1}^n g_i^t\right\rangle \tag{40}$$

$$= Q\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 + \mathbb{E}\left\langle\nabla f(w^t), g^t - Q\nabla f(w^t)\right\rangle \tag{41}$$

$$\overset{(32c)}{\geq} Q\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|g^t - Q\nabla f(w^t)\right\|^2 \tag{42}$$

$$= \frac{2Q-1}{2}\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|g^t - Q\nabla f(w^t)\right\|^2. \tag{43}$$

Moreover, the following holds for S_{a_2} in (37):

$$\mathbb{E}_{i_t}\left[S_{a_2}\right]=\mathbb{E}_{i_t}\left\|\sum_{q=0}^{Q-1}\tilde{\nabla}f_{i_t}\left(w_{i_t,q}^{\Omega(t)}\right)\right\|^2=\frac{1}{n}\sum_{i=1}^{n}\left\|\tilde{g}_i^t\right\|^2. \tag{44}$$

Now, according to (37), (43), and (44), we have:

$$\mathbb{E}f\left(w^{t+1}\right)\leq\mathbb{E}f(w^t)-\frac{\eta\beta(2Q-1)}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{45}$$

$$+\frac{\eta\beta}{2}\mathbb{E}\underbrace{\left\|g^t-Q\nabla f(w^t)\right\|^2}_{=:S_{a_3}}+\frac{L\eta^2\beta^2}{2n}\mathbb{E}\underbrace{\left[\sum_{i=1}^{n}\left\|\tilde{g}_i^t\right\|^2\right]}_{=:S_{a_4}}, \tag{46}$$

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

$$S_{a_3}=\left\|\frac{1}{n}\sum_{i=1}^{n}\left(g_i^t-Q\nabla f_i(w^t)\right)\right\|^2\overset{(32d)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\left\|g_i^t-Q\nabla f_i(w^t)\right\|^2 \tag{47}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{q=0}^{Q-1}\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-Q\nabla f_i(w^t)\right\|^2 \tag{48}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{q=0}^{Q-1}\left[\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i(w^t)\right]\right\|^2 \tag{49}$$

$$\overset{(32d)}{\leq}\frac{Q}{n}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i(w^t)\right\|^2, \tag{50}$$

$$S_{a_4}=\sum_{i=1}^{n}\left\|\sum_{q=0}^{Q-1}\tilde{\nabla}f_i\left(w_{i,q}^{\Omega(t)}\right)\right\|^2 \tag{51}$$

$$\overset{(32d)}{\leq}Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|\tilde{\nabla}f_i\left(w_{i,q}^{\Omega(t)}\right)\right\|^2 \tag{52}$$

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

$$\left.+\nabla f_i\left(w^t\right)-\nabla f\left(w^t\right)+\nabla f\left(w^t\right)\right\|^2 \tag{53}$$

$$\overset{(32d)}{\leq}4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left[\left\|\tilde{\nabla}f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)\right\|^2+\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i\left(w^t\right)\right\|^2\right.$$

$$\left.+\left\|\nabla f_i\left(w^t\right)-\nabla f\left(w^t\right)\right\|^2+\left\|\nabla f\left(w^t\right)\right\|^2\right]\Rightarrow \tag{54}$$

$$\mathbb{E}[S_{a_4}] \overset{(54)}{\leq} 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}_{p_i}\left[\left\|\tilde{\nabla} f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)\right\|^2\right] \tag{55}$$

$$+4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i\left(w^t\right)\right\|^2 \tag{56}$$

$$+4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i\left(w^t\right)-\nabla f\left(w^t\right)\right\|^2 \tag{57}$$

$$+4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f\left(w^t\right)\right\|^2 \tag{58}$$

$$\overset{(17),(18)}{\leq} 4nQ^2\left[\sigma_a^2+\gamma_g^2+\mathbb{E}\left\|\nabla f\left(w^t\right)\right\|^2\right] \tag{59}$$

$$+4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i(w^t)\right\|^2. \tag{60}$$

Therefore, due to (45)-(50) and (59)-(60), we have

$$\mathbb{E}f\left(w^{t+1}\right)\leq \mathbb{E}f(w^t)-\left[\frac{\eta\beta(2Q-1)}{2}-2\eta^2L\beta^2Q^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{61}$$

$$+\left[\frac{\eta\beta Q}{2n}+\frac{2\eta^2\beta^2QL}{n}\right]\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla f_i\left(w_{i,q}^{\Omega(t)}\right)-\nabla f_i(w^t)\right\|^2 \tag{62}$$

$$+2\eta^2L\beta^2Q^2\sigma_a^2+2\eta^2L\beta^2Q^2\gamma_g^2 \tag{63}$$

$$\overset{(14)}{\leq} \mathbb{E}f(w^t)-\left[\frac{\eta\beta(2Q-1)}{2}-2\eta^2L\beta^2Q^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{64}$$

$$+\frac{\eta\beta QL^2\left(1+4\eta\beta L\right)}{2n}\mathbb{E}\underbrace{\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|w_{i,q}^{\Omega(t)}-w^t\right\|^2}_{=:S_{a_5}} \tag{65}$$

$$+2\eta^2L\beta^2Q^2\sigma_a^2+2\eta^2L\beta^2Q^2\gamma_g^2. \tag{66}$$

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

$$\left\|w^t-w_{i,q}^{\Omega(t)}\right\|^2 \tag{67}$$

$$=\left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1}-w^s\right)+w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^2 \tag{68}$$

$$\overset{(32a)}{\leq}\left(1+\frac{1}{\beta^2}\right)\left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1}-w^s\right)\right\|^2+\left(1+\beta^2\right)\left\|w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^2 \tag{69}$$

$$\overset{(32d)}{\leq}\left(t-\Omega(t)\right)\left(1+\frac{1}{\beta^2}\right)\left[\sum_{s=\Omega(t)}^{t-1}\left\|w^{s+1}-w^s\right\|^2\right]+\left(1+\beta^2\right)\left\|w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^2 \tag{70}$$

$$\overset{(13)}{\leq}\tau\left(1+\frac{1}{\beta^2}\right)\underbrace{\left[\sum_{s=t-\tau}^{t-1}\left\|w^{s+1}-w^s\right\|^2\right]}_{=:S_{a_7}}+\left(1+\beta^2\right)\underbrace{\left\|w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^2}_{=:S_{a_6}}. \tag{71}$$

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:

$$\mathbb{E}\left\|w_{i,q}^s - w^s\right\|^2 = \mathbb{E}\left\|w_{i,q-1}^s - \eta\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) - w^s\right\|^2 \tag{72}$$

$$\begin{aligned} &= \mathbb{E}\Big\|w_{i,q-1}^s - w^s - \eta\nabla f\left(w^s\right) \\ &\quad - \eta\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) + \eta\nabla f_i\left(w_{i,q-1}^s\right) \\ &\quad - \eta\nabla f_i\left(w_{i,q-1}^s\right) + \eta\nabla f_i\left(w^s\right) \\ &\quad - \eta\nabla f_i\left(w^s\right) + \eta\nabla f\left(w^s\right)\Big\|^2 \end{aligned} \tag{73}$$

$$\overset{(32a)}{\leq} \left(1+\frac{1}{2Q}\right)\mathbb{E}\Big\|w_{i,q-1}^s - w^s\Big\|^2 \tag{74}$$

$$\begin{aligned} &+ \quad 4(1+2Q)\eta^2\mathbb{E}\Bigg[\Big\|\tilde{\nabla} f_i\left(w_{i,q-1}^s\right) - \nabla f_i\left(w_{i,q-1}^s\right)\Big\|^2 \\ &\qquad + \Big\|\nabla f_i\left(w_{i,q-1}^s\right) - \nabla f_i\left(w^s\right)\Big\|^2 \\ &\qquad + \Big\|\nabla f_i\left(w^s\right) - \nabla f\left(w^s\right)\Big\|^2 \\ &\qquad + \Big\|\nabla f\left(w^s\right)\Big\|^2\Bigg] \end{aligned} \tag{75}$$

$$\overset{(14),(17)}{\leq} \left(1+\frac{1}{2Q}\right)\mathbb{E}\Big\|w_{i,q-1}^s - w^s\Big\|^2 \tag{76}$$

$$\begin{aligned} &+ \quad 4(1+2Q)\eta^2\Bigg[\sigma_a^2 + L^2\,\mathbb{E}\Big\|w_{i,q-1}^s - w^s\Big\|^2 \\ &\qquad + \mathbb{E}\Big\|\nabla f_i\left(w^s\right) - \nabla f\left(w^s\right)\Big\|^2 + \mathbb{E}\Big\|\nabla f\left(w^s\right)\Big\|^2\Bigg]. \end{aligned} \tag{77}$$

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

$$\eta^2 \leq \frac{1}{16L^2(Q+1)^2} \leq \frac{1}{8L^2Q(2Q+1)} \Rightarrow 4(1+2Q)\eta^2L^2 \leq \frac{1}{2Q}, \tag{78}$$

therefore, due to (72)-(78), we have:

$$\underbrace{\mathbb{E}\left\|w_{i,q}^{s}-w^{s}\right\|^{2}}_{:=P_{i,q}^{s}} \leq \underbrace{\left(1+\frac{1}{Q}\right)\mathbb{E}\left\|w_{i,q-1}^{s}-w^{s}\right\|^{2}}_{:=P_{i,q-1}^{s}} \tag{79}$$

$$+\underbrace{4(1+2Q)\eta^{2}\left[\sigma_{a}^{2}+\mathbb{E}\left\|\nabla f_{i}\left(w^{s}\right)-\nabla f\left(w^{s}\right)\right\|^{2}+\mathbb{E}\left\|\nabla f\left(w^{s}\right)\right\|^{2}\right]}_{:=R_{i}^{s}} \Rightarrow \tag{80}$$

$$P_{i,q}^{s} \leq \left(1+\frac{1}{Q}\right)P_{i,q-1}^{s}+R_{i}^{s} \tag{81}$$

$$=R_{i}^{s}\sum_{k=0}^{q-1}\left(1+\frac{1}{Q}\right)^{k} \leq R_{i}^{s}\sum_{k=0}^{Q-1}\left(1+\frac{1}{Q}\right)^{k} \tag{82}$$

$$=R_{i}^{s}\frac{\left(1+\frac{1}{Q}\right)^{Q}-1}{\left(1+\frac{1}{Q}\right)-1}=R_{i}^{s}Q\left[\left(1+\frac{1}{Q}\right)^{Q}-1\right] \leq R_{i}^{s}Q(e-1) \leq 2R_{i}^{s}Q \Rightarrow \tag{83}$$

$$\mathbb{E}\left\|w_{i,q}^{s}-w^{s}\right\|^{2} \leq 8Q(1+2Q)\eta^{2}\left[\sigma_{a}^{2}+\mathbb{E}\left\|\nabla f_{i}\left(w^{s}\right)-\nabla f\left(w^{s}\right)\right\|^{2}+\mathbb{E}\left\|\nabla f\left(w^{s}\right)\right\|^{2}\right], \tag{84}$$

for all $q \in [Q]$ and $s \geq 0$. We now will use (72)-(84) to provide a bound on the expression in S_{a_7}. Again, note that according to Algorithms 1 & 3, we have:

$$w^{s+1}=w^{s}-\beta\left(w_{i_s,0}^{\Omega(s)}-w_{i_s,Q}^{\Omega(s)}\right) \Rightarrow \tag{85}$$

$$\mathbb{E}\left\|w^{s+1}-w^{s}\right\|^{2} \leq \beta^{2}\,\mathbb{E}\left\|w_{i_s,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2} \tag{86}$$

$$=\beta^{2}\,\mathbb{E}\left[\mathbb{E}_{i_s}\left\|w_{i_s,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2}\right] \tag{87}$$

$$=\frac{\beta^{2}}{n}\sum_{j=1}^{n}\mathbb{E}\left\|w_{j,Q}^{\Omega(s)}-w^{\Omega(s)}\right\|^{2} \tag{88}$$

$$\leq 8Q(1+2Q)\eta^{2}\beta^{2}\left[\sigma_{a}^{2}+\gamma_{g}^{2}+\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^{2}\right]. \tag{89}$$

Let $\phi=8\eta^{2}Q^{2}(1+2Q)(1+\beta^{2})$, then according to (67)-(89)

$$\frac{1}{n\phi}\mathbb{E}[S_{a_5}] \leq \tau\left[\sum_{s=t-\tau}^{t-1}\left\|w^{s+1}-w^{s}\right\|^{2}\right]+\frac{1}{nQ}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|w^{\Omega(t)}-w_{i,q}^{\Omega(t)}\right\|^{2}. \tag{90}$$

$$\leq \tau^{2}\sigma_{a}^{2}+\tau^{2}\gamma_{g}^{2}+\tau\sum_{s=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^{2} \tag{91}$$

$$+\sigma_{a}^{2}+\gamma_{g}^{2}+\mathbb{E}\left\|\nabla f\left(w^{\Omega(t)}\right)\right\|^{2} \tag{92}$$

$$\leq (\tau^{2}+1)\left[\sigma_{a}^{2}+\gamma_{g}^{2}\right]+\mathbb{E}\left\|\nabla f\left(w^{\Omega(t)}\right)\right\|^{2}+\tau\sum_{s=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^{2} \tag{93}$$

$$\leq (\tau^{2}+1)\left[\sigma_{a}^{2}+\gamma_{g}^{2}\right]+\tau\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^{2}. \tag{94}$$

Thus, by combining (61)-(94), we have the following inequality:

$$\mathbb{E}f\left(w^{t+1}\right) \leq \mathbb{E}f(w^t) - \eta\beta\left[\frac{2Q-1}{2} - 2\eta\beta LQ^2\right]\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{95}$$

$$+ 4\eta^3\beta L^2Q^3(1{+}2Q)(1{+}\beta^2)(1{+}4\eta\beta L)\,\tau\left[\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2\right] \tag{96}$$

$$+ 4\eta^3\beta L^2Q^3(1{+}2Q)(\tau^2{+}1)(1{+}\beta^2)(1{+}4\eta\beta L)\left(\sigma_a^2+\gamma_g^2\right) \tag{97}$$

$$+ 2\eta^2\beta^2LQ^2\left(\sigma_a^2+\gamma_g^2\right), \tag{98}$$

where by rearranging, we obtain the following inequality:

$$(1-4\eta\beta LQ)\,\mathbb{E}\left\|\nabla f(w^t)\right\|^2 \tag{99}$$

$$- 8\eta^2L^2Q^2(1{+}2Q)(1{+}\beta^2)(1{+}4\eta\beta L)\,\tau\left[\sum_{s=t-\tau}^{t}\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2\right] \tag{100}$$

$$\leq \frac{2\left[\mathbb{E}f(w^t)-\mathbb{E}f\left(w^{t+1}\right)\right]}{\eta\beta Q} \tag{101}$$

$$+ 8\eta^2L^2Q^2(1{+}2Q)(\tau^2{+}1)(1{+}\beta^2)(1{+}4\eta\beta L)\left(\sigma_a^2+\gamma_g^2\right) \tag{102}$$

$$+ 4\eta\beta LQ\left(\sigma_a^2+\gamma_g^2\right). \tag{103}$$

Now, note that for any $s \geq 0$,[2]

$$\mathbb{E}\left\|\nabla f\left(w^{\Omega(s)}\right)\right\|^2 \leq \sum_{u=s-\tau}^{s}\mathbb{E}\left\|\nabla f\left(w^u\right)\right\|^2, \tag{104}$$

Therefore, we add up the inequality in (99)-(103), for $t = 0, 1, \ldots T{-}1$, and obtain

$$\left[1-4\eta\beta LQ-8\eta^2L^2Q^2(1{+}2Q)\tau(\tau{+}1)^2(1{+}\beta^2)(1{+}4\eta\beta L)\right]\frac{\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(w^t)\right\|^2}{T} \tag{105}$$

$$\leq \frac{2\left[f(w^0)-\mathbb{E}f(w^T)\right]}{\eta\beta QT} + 4\eta\beta LQ\left(\sigma_a^2+\gamma_g^2\right) \tag{106}$$

$$+ 8\eta^2L^2Q^2(1{+}2Q)(\tau^2{+}1)(1{+}\beta^2)(1{+}4\eta\beta L)\left(\sigma_a^2+\gamma_g^2\right). \tag{107}$$

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

$$1-4\eta\beta LQ-8\eta^2L^2Q^2(1{+}2Q)\tau(\tau{+}1)^2(1{+}\beta^2)(1{+}4\eta\beta L) \geq \frac{1}{2}, \tag{108}$$

$$\eta \leq \frac{1}{4L(Q{+}1)}, \tag{109}$$

for $T \geq 160L(Q{+}7)(\tau{+}1)^3$. Therefore, we can conclude the final result in Theorem 2 under this choice of η and β. □

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

D Personalized Asynchronous Federated Learning: MAML

Proof of Theorem 3. To simplify (11), 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 (36), at round t, the update rule for Option B can be written as follows:

$$w^{t+1} = w^t - \eta\beta \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right). \tag{110}$$

According to Lemma 1,

$$F^{(b)}\left(w^{t+1}\right) \overset{(14)}{\leq} F^{(b)}(w^t) - \eta\beta \underbrace{\left\langle \nabla F^{(b)}(w^t), \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right) \right\rangle}_{=:S_{b_1}} + \frac{L_b \eta^2 \beta^2}{2} \underbrace{\left\| \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right) \right\|^2}_{=:S_{b_2}} \tag{111}$$

Similar to the inequalities in (38)-(43), we first show a lower bound on term S_{b_1} in (111). 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

$$\left\| \mathbb{E}\left[\tilde{g}^t - g^t \right] \right\| \overset{(32b)}{\leq} \frac{1}{n} \sum_{i=1}^n \left\| \mathbb{E}\left[\tilde{g}_i^t - g_i^t \right] \right\| \tag{112}$$

$$\overset{(32b)}{\leq} \frac{1}{n} \sum_{i=1}^n \sum_{q=0}^{Q-1} \left\| \mathbb{E}\left[\tilde{\nabla} F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) - \nabla F_i^{(b)} \left(w_{i,q}^{\Omega(t)} \right) \right] \right\| \tag{113}$$

$$\overset{(24)}{\leq} \frac{1}{n} \sum_{i=1}^n \sum_{q=0}^{Q-1} \mu_b = Q\mu_b, \tag{114}$$

where as we discussed in (24), μ_b measures the unbiasedness in the estimation of the personalized stochastic gradient.

$$\mathbb{E}\left[S_{b_1}\right] = \mathbb{E}\left[\mathbb{E}_{i_t} \left\langle \nabla F^{(b)}(w^t), \tilde{g}_{i_t}^t \right\rangle \right] \tag{115}$$

$$= \mathbb{E}\left[\left\langle \nabla F^{(b)}(w^t), \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^t \right\rangle \right] = \mathbb{E}\left[\left\langle \nabla F^{(b)}(w^t), \tilde{g}^t \right\rangle \right] \tag{116}$$

$$= Q\, \mathbb{E} \left\| \nabla F^{(b)}(w^t) \right\|^2 + \mathbb{E} \left\langle \nabla F^{(b)}(w^t), \mathbb{E}\left[\tilde{g}^t - g^t \right] \right\rangle \tag{117}$$

$$+ \mathbb{E} \left\langle \nabla F^{(b)}(w^t), g^t - Q \nabla F^{(b)}(w^t) \right\rangle \tag{118}$$

$$\overset{(32c)}{\geq} Q\, \mathbb{E} \left\| \nabla F^{(b)}(w^t) \right\|^2 - \frac{1}{4} \mathbb{E} \left\| \nabla F^{(b)}(w^t) \right\|^2 - \left\| \mathbb{E}\left[g^t - \tilde{g}^t \right] \right\|^2 \tag{119}$$

$$- \frac{1}{4} \mathbb{E} \left\| \nabla F^{(b)}(w^t) \right\|^2 - \mathbb{E} \left\| g^t - Q \nabla F^{(b)}(w^t) \right\|^2 \tag{120}$$

$$\overset{(112)-(114)}{\geq} \frac{2Q-1}{2} \mathbb{E} \left\| \nabla F^{(b)}(w^t) \right\|^2 - \mathbb{E} \left\| g^t - Q \nabla F^{(b)}(w^t) \right\|^2 - Q^2 \mu_b^2, \tag{121}$$

and

$$\mathbb{E}_{i_t}\left[S_{b_2}\right] = \mathbb{E}_{i_t} \left\| \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(b)} \left(w_{i_t,q}^{\Omega(t)} \right) \right\|^2 = \frac{1}{n} \sum_{i=1}^n \left\| \tilde{g}_i^t \right\|^2. \tag{122}$$

Therefore, according to (111), (121), and (122),

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \frac{\eta\beta(2Q-1)}{2}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 + \eta\beta Q^2\mu_b^2 \tag{123}$$

$$+ \eta\beta\,\underbrace{\mathbb{E}\left\|g^t - Q\nabla F^{(b)}(w^t)\right\|^2}_{=:S_{b_3}} + \frac{L_b\eta^2\beta^2}{2n}\underbrace{\mathbb{E}\sum_{i=1}^{n}\left\|\tilde{g}_i^t\right\|^2}_{=:S_{b_4}}, \tag{124}$$

where similar to (47)-(50), we can bound S_{b_3} as follows:

$$S_{b_3} \leq \frac{Q}{n}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}(w^t)\right\|^2. \tag{125}$$

Moreover, we can show an upper bound on S_{b_4} akin to (126)-(129):

$$S_{b_4} = \sum_{i=1}^{n}\left\|\sum_{q=0}^{Q-1}\tilde{\nabla}F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right)\right\|^2 \tag{126}$$

$$\overset{(32d)}{\leq} Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right)\right\|^2 \tag{127}$$

$$= Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) + \nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w^t\right)\right.$$
$$\left. + \nabla F_i^{(b)}\left(w^t\right) - \nabla F^{(b)}\left(w^t\right) + \nabla F^{(b)}\left(w^t\right)\right\|^2 \tag{128}$$

$$\overset{(32d)}{\leq} 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\left[\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right)\right\|^2\right.$$
$$+ \left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w^t\right)\right\|^2$$
$$+ \left\|\nabla F_i^{(b)}\left(w^t\right) - \nabla F^{(b)}\left(w^t\right)\right\|^2$$
$$\left. + \left\|\nabla F^{(b)}\left(w^t\right)\right\|^2\right] \Rightarrow \tag{129}$$

$$\mathbb{E}[S_{b_4}] \overset{(129)}{\leq} 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}_{p_i}\left[\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right)\right\|^2\right] \tag{130}$$

$$+ 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w^t\right)\right\|^2 \tag{131}$$

$$+ 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_i^{(b)}\left(w^t\right) - \nabla F^{(b)}\left(w^t\right)\right\|^2 \tag{132}$$

$$+ 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F^{(b)}\left(w^t\right)\right\|^2 \tag{133}$$

$$\overset{(17),(18)}{\leq} 4nQ^2\left[\sigma_b^2 + \gamma_b^2 + \mathbb{E}\left\|\nabla F^{(b)}\left(w^t\right)\right\|^2\right] \tag{134}$$

$$+ 4Q\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}(w^t)\right\|^2. \tag{135}$$

Therefore, due to (123)-(125) and (134)-(135), we have

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_b \beta^2 Q^2\right] \mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{136}$$

$$+ \left[\frac{\eta\beta Q}{n} + \frac{2\eta^2\beta^2 Q L_b}{n}\right] \sum_{i=1}^{n}\sum_{q=0}^{Q-1} \mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,q}^{\Omega(t)}\right) - \nabla F_i^{(b)}(w^t)\right\|^2 \tag{137}$$

$$+ \eta\beta Q^2\mu_b^2 + 2\eta^2 L_b\beta^2 Q^2\sigma_b^2 + 2\eta^2 L_b\beta^2 Q^2\gamma_b^2 \tag{138}$$

$$\overset{(14)}{\leq} \mathbb{E}F^{(b)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_b \beta^2 Q^2\right] \mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{139}$$

$$+ \frac{\eta\beta Q L_b^2\left(1+2\eta\beta L_b\right)}{n} \sum_{i=1}^{n}\sum_{q=0}^{Q-1} \mathbb{E}\underbrace{\left\|w_{i,q}^{\Omega(t)} - w^t\right\|^2}_{=:S_{b_5}} \tag{140}$$

$$+ \eta\beta Q^2\mu_b^2 + 2\eta^2 L_b\beta^2 Q^2\left(\sigma_b^2 + \gamma_b^2\right). \tag{141}$$

Now, we provide an upper bound on S_{b_5} in (139) as follows:

$$S_{b_5} = \left\|w^t - w_{i,q}^{\Omega(t)}\right\|^2 = \left\|w^t - w^{\Omega(t)} + w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{142}$$

$$\overset{(32a)}{\leq} 2\underbrace{\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2}_{S_{b_6}} + 2\underbrace{\left\|w^t - w^{\Omega(t)}\right\|^2}_{S_{b_7}}, \tag{143}$$

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

$$\mathbb{E}[S_{b_6}] = \mathbb{E}\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{144}$$

$$= \mathbb{E}\left\|w_{i,0}^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2 \tag{145}$$

$$\overset{(110)}{=} \eta^2\mathbb{E}\left\|\sum_{r=0}^{q-1}\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 \tag{146}$$

$$\overset{(32d)}{\leq} \eta^2 q\sum_{r=0}^{q-1}\mathbb{E}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 \tag{147}$$

$$\overset{(32d)}{\leq} 2\eta^2 q\sum_{r=0}^{q-1}\left[\mathbb{E}\left\|\tilde{\nabla}F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right) - \nabla F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2 + \mathbb{E}\left\|\nabla F_i^{(b)}\left(w_{i,r}^{\Omega(t)}\right)\right\|^2\right] \tag{148}$$

$$\overset{(25),(27)}{\leq} 2\eta^2 q\sum_{r=0}^{q-1}\left(G_b^2 + \sigma_b^2\right) = 2\eta^2 q^2\left(G_b^2 + \sigma_b^2\right), \tag{149}$$

$$\mathbb{E}[S_{b_7}] = \mathbb{E}\left\|w^t - w^{\Omega(t)}\right\|^2 \tag{150}$$

$$= \mathbb{E}\left\|\sum_{s=\Omega(t)}^{t-1}\left(w^{s+1}-w^s\right)\right\|^2 \tag{151}$$

$$\overset{\text{Alg. 1,3}}{=} \eta^2\beta^2\mathbb{E}\left\|\sum_{s=\Omega(t)}^{t-1}\sum_{q=0}^{Q-1}\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2 \tag{152}$$

$$\overset{\text{(32d)}}{\leq} \eta^2\beta^2 Q\left(t-\Omega(t)\right)\sum_{s=\Omega(t)}^{t-1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2 \tag{153}$$

$$\overset{(13)}{\leq} 2\eta^2\beta^2 Q\,\tau\sum_{s=t-\tau}^{t-1}\sum_{q=0}^{Q-1}\left[\mathbb{E}\left\|\tilde{\nabla}F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)-\nabla F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2\right.$$

$$\left.+\,\mathbb{E}\left\|\nabla F_{i_s}^{(b)}\left(w_{i_s,q}^{\Omega(s)}\right)\right\|^2\right] \tag{154}$$

$$\overset{\text{(25),(27)}}{\leq} 2\eta^2\beta^2 Q\tau^2\sum_{q=0}^{Q-1}\left(G_b^2+\sigma_b^2\right) = 2\eta^2\beta^2Q^2\tau^2\left(G_b^2+\sigma_b^2\right). \tag{155}$$

So, according to (136)-(155),

$$\mathbb{E}F^{(b)}\left(w^{t+1}\right) \leq \mathbb{E}F^{(b)}(w^t) - \frac{\eta\beta}{2}\left(2Q-1-4\eta\beta L_bQ^2\right)\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \tag{156}$$

$$+\,4\eta^3\beta Q^4L_b^2\left(1+2\eta\beta L_bQ\right)\left(G_b^2+\sigma_b^2\right)\left(\beta^2\tau^2+1\right) \tag{157}$$

$$+\,\eta\beta Q^2\mu_b^2 + 2\eta^2\beta^2L_bQ^2\sigma_b^2 + 2\eta^2\beta^2L_bQ^2\gamma_b^2, \tag{158}$$

where by adding the terms in (156)-(158), for $t = 0, 1, \ldots T-1$, and rearranging them, we obtain the following inequality:

$$\frac{1-4\eta\beta L_bQ}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F^{(b)}(w^t)\right\|^2 \leq \frac{2\left(F^{(b)}(w^0)-\mathbb{E}F^{(b)}(w^T)\right)}{\eta\beta QT} \tag{159}$$

$$+\,8\eta^2Q^3L_b^2\left(1+2\eta\beta L_bQ\right)\left(G_b^2+\sigma_b^2\right)\left(\beta^2\tau^2+1\right) \tag{160}$$

$$+\,4\eta\beta L_bQ\left(\sigma_b^2+\gamma_b^2\right) \tag{161}$$

$$+\,2Q\mu_b^2. \tag{162}$$

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 $\eta \leq \frac{1}{8\beta L_bQ}$. □

E Personalized Asynchronous Federated Learning: ME

We start by showing (12), where according to the definitions in (8) and (10), we have

$$\hat{\theta}_i(w) = \underset{\theta_i \in \mathbb{R}^d}{\arg\min} \left[f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - w\|^2 \right] \Rightarrow \nabla f_i \left(\hat{\theta}_i(w) \right) + \lambda \left[\hat{\theta}_i(w) - w \right] = 0, \tag{163}$$

$$F_i^{(c)}(w) = f_i \left(\hat{\theta}_i(w) \right) + \frac{\lambda}{2} \left\| \hat{\theta}_i(w) - w \right\|^2, \tag{164}$$

therefore,

$$\nabla F_i^{(c)}(w) \overset{(164)}{=} \frac{\partial \hat{\theta}_i(w)}{\partial w} \left[\nabla f_i \left(\hat{\theta}_i(w) \right) \right] + \lambda \left[\frac{\partial \hat{\theta}_i(w)}{\partial w} - I \right] \left[\hat{\theta}_i(w) - w \right] \tag{165}$$

$$\overset{(163)}{=} \lambda \frac{\partial \hat{\theta}_i(w)}{\partial w} \left[w - \hat{\theta}_i(w) \right] + \lambda \left[\frac{\partial \hat{\theta}_i(w)}{\partial w} - I \right] \left[\hat{\theta}_i(w) - w \right] \tag{166}$$

$$= \lambda \left[w - \hat{\theta}_i(w) \right]. \tag{167}$$

Before, presenting the proof of Theorem 4, 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:

$$\nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \overset{(167)}{=} \lambda \left[w - \hat{\theta}_i(w) \right] - \lambda \left[v - \hat{\theta}_i(v) \right] \tag{168}$$

$$\overset{(163)}{=} \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \Rightarrow \tag{169}$$

$$\left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| = \left\| \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \tag{170}$$

$$\overset{(14)}{\leq} L \left\| \hat{\theta}_i(w) - \hat{\theta}_i(v) \right\| \tag{171}$$

$$\overset{(163)}{=} L \left\| w - \frac{1}{\lambda} \nabla f_i \left(\hat{\theta}_i(w) \right) - v + \frac{1}{\lambda} \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \tag{172}$$

$$\leq L \|w - v\| + \frac{L}{\lambda} \left\| \nabla f_i \left(\hat{\theta}_i(w) \right) - \nabla f_i \left(\hat{\theta}_i(v) \right) \right\| \tag{173}$$

$$= L \|w - v\| + \frac{L}{\lambda} \left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| \Rightarrow \tag{174}$$

$$\left\| \nabla F_i^{(c)}(w) - \nabla F_i^{(c)}(y) \right\| \leq \frac{\lambda L}{\lambda - L} \|w - v\|, \tag{175}$$

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

$$\frac{\lambda L}{\lambda - L} \leq L_c \coloneqq \frac{\lambda}{\kappa - 1} \tag{176}$$

This concludes the statement of Lemma 5. □

Proof of Lemma 6. According to Step 11 of Algorithm 3, let us introduce full and stochastic auxiliary cost functions $h_i(\cdot)$ and $\tilde{h}_i(\cdot)$ as follows:

$$h_i(\theta_i, w) = f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - w\|^2, \tag{177}$$

$$\tilde{h}_i(\theta_i, w, \mathcal{D}) = \tilde{f}_i(\theta_i, \mathcal{D}) + \frac{\lambda}{2} \|\theta_i - w\|^2, \tag{178}$$

where due to (177), we have

$$\nabla \tilde{h}_i(\tilde{\theta}_i(w), w, \mathcal{D}) = \nabla \tilde{f}_i(\tilde{\theta}_i(w), \mathcal{D}) + \lambda \left[\tilde{\theta}_i(w) - w \right], \tag{179}$$

hence, we can show (28) as follows:

$$\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w,\mathcal{D})-\nabla F_i^{(c)}(w)\right]\right\| \tag{180}$$

$$\overset{(9),(12)}{=}\left\|\mathbb{E}_{p_i}\left[\lambda\hat{\theta}_i(w)-\lambda\tilde{\theta}_i(w)\right]\right\| \tag{181}$$

$$\overset{(163),(178)}{=}\left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w))-\nabla\tilde{f}_i(\tilde{\theta}_i(w),\mathcal{D})+\nabla\tilde{h}_i(\tilde{\theta}_i(w),w,\mathcal{D})\right]\right\| \tag{182}$$

$$=\left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w))-\nabla f_i(\tilde{\theta}_i(w))\right]+\mathbb{E}_{p_i}\left[\nabla\tilde{h}_i(\tilde{\theta}_i(w),w,\mathcal{D})\right]\right\| \tag{183}$$

$$\leq\left\|\mathbb{E}_{p_i}\left[\nabla f_i(\hat{\theta}_i(w))-\nabla f_i(\tilde{\theta}_i(w))\right]\right\|+\nu \tag{184}$$

$$\overset{(14)}{\leq} L\left\|\mathbb{E}_{p_i}\left[\hat{\theta}_i(w)-\tilde{\theta}_i(w)\right]\right\|+\nu \tag{185}$$

$$\overset{(181)}{=}\frac{L}{\lambda}\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w,\mathcal{D})-\nabla F_i^{(c)}(w)\right]\right\|+\nu\Rightarrow \tag{186}$$

$$\left\|\mathbb{E}_{p_i}\left[\nabla\tilde{F}_i^{(c)}(w,\mathcal{D})-\nabla F_i^{(c)}(w)\right]\right\|\leq\frac{\lambda}{\lambda-L}\nu. \tag{187}$$

You can find the proof of (29) in [6][Appendix A.2]. □

Proof of Lemma 7. First, note that we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla F_i^{(c)}(w)-\nabla F^{(c)}(w)\right\|^2 \tag{188}$$

$$\overset{(167)}{=}\frac{1}{n}\sum_{i=1}^{n}\left\|\lambda(w-\hat{\theta}_i(w))-\frac{1}{n}\sum_{j=1}^{n}\lambda(w-\hat{\theta}_i(w))\right\|^2 \tag{189}$$

$$\overset{(163)}{=}\frac{1}{n^3}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}\left[\nabla f_i(\hat{\theta}_i(w))-\nabla f_j(\hat{\theta}_j(w))\right]\right\|^2 \tag{190}$$

$$\overset{(32d)}{\leq}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\nabla f_i(\hat{\theta}_i(w))-\nabla f_j(\hat{\theta}_j(w))\right\|^2. \tag{191}$$

So, we simplify the upper bound as follows:

$$\left\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\right\|^2 \tag{192}$$

$$= \Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w)) + \nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w)) + \nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w)) + \nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w)) + \nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{193}$$

$$\overset{(32a)}{\leq} \frac{4}{3}\Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w)) + \nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w)) + \nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{194}$$

$$+ 4\Big\|\nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w)) + \nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \tag{195}$$

$$\overset{(32d)}{\leq} 4\Big\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w))\Big\|^2 \tag{196}$$

$$+ 4\Big\|\nabla f(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_i(w))\Big\|^2 \tag{197}$$

$$+ 4\Big\|\nabla f_j(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_j(w))\Big\|^2 \tag{198}$$

$$+ 8\Big\|\nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \tag{199}$$

$$+ 8\Big\|\nabla f(\hat{\theta}_i(w)) - \nabla f_j(\hat{\theta}_i(w))\Big\|^2 \tag{200}$$

Note that we can bound (199) and (200) according to Lemma 18:

$$\frac{1}{n}\sum_{i=1}^{n}\Big\|\nabla f_i(\hat{\theta}_j(w)) - \nabla f(\hat{\theta}_j(w))\Big\|^2 \overset{(18)}{\leq} \gamma_g^2, \tag{201}$$

and also given the fact that function $f(\cdot)$ as well as each function $f_i(\cdot)$ are L-smooth, we can bound (196), (197), and (198) as follows:

$$\left\|\nabla f_i(\hat{\theta}_i(w)) - \nabla f_i(\hat{\theta}_j(w))\right\|^2 \tag{202}$$

$$\leq L^2\Big\|\hat{\theta}_i(w) - \hat{\theta}_j(w)\Big\|^2 \tag{203}$$

$$= \frac{L^2}{\lambda^2}\Big\|\lambda\left[\hat{\theta}_i(w) - w\right] - \lambda\left[\hat{\theta}_j(w) - w\right]\Big\|^2 \tag{204}$$

$$\overset{(167)}{=} \frac{L^2}{\lambda^2}\Big\|\nabla F_i^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2 \tag{205}$$

$$= \frac{L^2}{\lambda^2}\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w) + \nabla F^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2 \tag{206}$$

$$\overset{(32d)}{\leq} \frac{2L^2}{\lambda^2}\left[\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\Big\|^2 + \Big\|\nabla F^{(c)}(w) - \nabla F_j^{(c)}(w)\Big\|^2\right]. \tag{207}$$

Therefore, according to (208)-(207), we have

$$\frac{1}{n}\sum_{i=1}^{n}\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\Big\|^2 \leq 16\gamma_g^2 + \frac{48L^2}{n\lambda^2}\sum_{i=1}^{n}\Big\|\nabla F_i^{(c)}(w) - \nabla \quad F^{(c)}(w)\Big\|^2 \Rightarrow \tag{208}$$

$$\frac{1}{n}\sum_{i=1}^{n}\Big\|\nabla F_i^{(c)}(w) - \nabla F^{(c)}(w)\Big\|^2 \leq \frac{16\lambda^2\gamma_g^2}{\lambda^2 - 48L^2}, \tag{209}$$

which concludes the proof. □

Now, we are ready to state the proof of Theorem 4.

Proof of Theorem 4. We write $\tilde{\nabla} F_i^{(c)}(w) = \nabla \tilde{F}_i^{(c)}(w, \mathcal{D}_i)$ to simplify (12). Then, the update rule for Algorithms 1 & 3 under Option C can be written as follows:

$$w^{t+1} = w^t - \eta\beta \sum_{q=0}^{Q-1} \tilde{\nabla} F_{i_t}^{(c)} \left(w_{i_t,q}^{\Omega(t)} \right), \tag{210}$$

where similar to (111)-(141), we can show that:

$$\mathbb{E} F^{(c)}\left(w^{t+1}\right) \leq \mathbb{E} F^{(c)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_c \beta^2 Q^2\right] \mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2 \tag{211}$$

$$+ \frac{\eta\beta Q L_c^2 (1+2\eta\beta L_c)}{n} \underbrace{\sum_{i=1}^{n} \sum_{q=0}^{Q-1} \mathbb{E}\left\|w_{i,q}^{\Omega(t)} - w^t\right\|^2}_{=:S_{c_1}} \tag{212}$$

$$+ \eta\beta Q^2 \mu_c^2 + 2\eta^2 L_c \beta^2 Q^2 \left(\sigma_c^2 + \gamma_c^2\right), \tag{213}$$

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 (8), it would only be sufficient to provide an upper bound on S_{c_1}. First, note that similar to (67)-(71), we have

$$\left\|w_{i,q}^{\Omega(t)} - w^t\right\|^2 \leq \tau \left(1 + \frac{1}{\beta^2}\right) \left[\sum_{s=t-\tau}^{t-1} \underbrace{\left\|w^{s+1} - w^s\right\|^2}_{=:S_{c_3}}\right] + \left(1+\beta^2\right) \underbrace{\left\|w^{\Omega(t)} - w_{i,q}^{\Omega(t)}\right\|^2}_{=:S_{c_2}}. \tag{214}$$

Now, if we introduce stepsize η such that $\eta \leq \frac{1}{4L_c(Q+1)}$, similar to (72)-(77) and (85)-(89), the following two inequalities holds for S_{c_2} and S_{c_3}:

$$\mathbb{E}[S_{c_2}] = \mathbb{E}\left\|w_{i,q}^{\Omega(t)} - w^{\Omega(t)}\right\|^2 \tag{215}$$

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

$$\mathbb{E}\left[S_{c_3}\right] = \mathbb{E}\left\|w^{s+1} - w^s\right\|^2 \leq 8Q(1+2Q)\eta^2\beta^2 \left[\sigma_c^2 + \gamma_c^2 + \mathbb{E}\left\|\nabla F^{(c)}\left(w^{\Omega(s)}\right)\right\|^2\right], \tag{216}$$

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

$$\frac{1}{n\phi}\mathbb{E}[S_{c_1}] \leq (\tau^2+1)\left[\sigma_c^2 + \gamma_c^2\right] + \tau \sum_{s=t-\tau}^{t} \sum_{u=s-\tau}^{s} \mathbb{E}\left\|\nabla F^{(c)}\left(w^u\right)\right\|^2. \tag{217}$$

Then, according to (211)-(217), we obtain

$$\mathbb{E} F^{(c)}(w^{t+1}) \leq \mathbb{E} F^{(c)}(w^t) - \left[\frac{\eta\beta(2Q-1)}{2} - 2\eta^2 L_c \beta^2 Q^2\right] \mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2 \tag{218}$$

$$+ 8\eta^3\beta Q^3 L_c^2 (1+2Q)(1+\beta^2)\left(1+2\eta\beta L_c\right)\tau \left[\sum_{s=t-\tau}^{t} \sum_{u=s-\tau}^{s} \mathbb{E}\left\|\nabla F^{(c)}(w^u)\right\|^2\right] \tag{219}$$

$$+ 8\eta^3\beta Q^3 L_c^2 (1+2Q)(\tau^2+1)(1+\beta^2)\left(1+2\eta\beta L_c\right)\left(\sigma_c^2 + \gamma_c^2\right) \tag{220}$$

$$+ 2\eta^2 L_c \beta^2 Q^2 \left(\sigma_c^2 + \gamma_c^2\right) \tag{221}$$

$$+ \eta\beta Q^2 \mu_c^2, \tag{222}$$

where by averaging the terms in (218)-(222), for $t = 0, 1, \dots T-1$, and rearranging them (similar to (105)-(107), we can conclude the following inequality:

$$\frac{1 - 4\eta\beta Q L_c - 16\eta^2 Q^2 L_c^2 (1+2Q)\tau(\tau+1)^2(1+\beta^2)\,(1+2\eta\beta L_c)}{T} \sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla F^{(c)}(w^t)\right\|^2$$

$$\leq \frac{2\left(F^{(c)}(w^0) - \mathbb{E}F^{(c)}(w^T)\right)}{\eta\beta Q T} + 2Q\mu_c^2 \tag{223}$$

$$+ 16\eta^2 Q^2 L_c^2 (1+2Q)(\tau^2+1)(1+\beta^2)\,(1+2\eta\beta L_c)\left(\sigma_c^2 + \gamma_c^2\right) \tag{224}$$

$$+ 4\eta\beta Q L_c \left(\sigma_c^2 + \gamma_c^2\right) \tag{225}$$

Finally, by fixing $\eta = \frac{1}{Q\sqrt{L_c T}}$, for $T \geq 288 L_c (Q+7)(\tau+1)^2$, we obtain the sublinear convergence rate in Theorem 4. □

F Numerical Experiments

In this section, we discuss the concurrency, speed-up, and accuracy of our proposed method in heterogeneous settings under delayed communications.

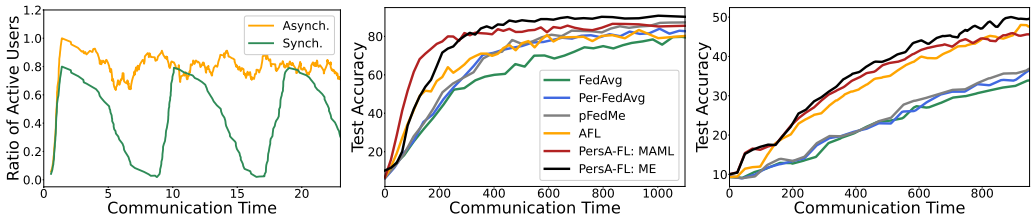
Let us first start by explaining our simulation setup for communications with delays. We consider a set of $n = 50$ different clients, where each client holds a set of random delays at the upload and download stage. We set the random delays to be generated such that the upload delay is 4 to 6 times higher than the download delay on average. We assume that the time for local updates is negligible compared to the time for communication and aggregation. Then, we plot the number of active users (not idle) over the training process under asynchronous communications. The orange chart in Figure 2**(a)** describes the ratio of active users, which on average is roughly 80% over time. Therefore, we consider the same ratio for client sampling in the synchronous updates for methods such as `FedAvg` and plot the activity of users in the same figure with green color. As shown in Figure 2**(a)**, the degree of concurrency for asynchronous methods is dramatically higher than their synchronous counterparts.

We also generate heterogeneous data from MNIST [21] and CIFAR-10 [20] datasets on the clients, i.e., each client holds a distinguished and unbalanced distribution of images from different classes. For instance, client $i \in [n]$ may mostly have samples from odd numbers in MNIST, while client $j \in [n]$ mainly holds digits zero to four. Over the underlying communication setup and heterogeneous data setting, we compare the speed and accuracy of `FedAvg`, `Per-FedAvg`, `pFedMe`, `AFL`, `PersA-FL: MAML`, `PersA-FL: ME`, where the first three methods are synchronous and the rest are asynchronous. For MNIST and CIFAR-10, we consider 2-layer and 4-layer convolutional networks [20] with pooling and dropout as well as cross-entropy loss. For all algorithms, we consider $Q = 10$ local updates, and select the best $\lambda \in \{15, 20, 25, 30\}$ for `ME` and $\alpha \in \{0.001, 0.002, 0.005, 0.01, 0.02\}$ for `MAML`. Moreover, we pick $\beta \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$ and fix $\eta = 0.03$. For both experiments, we consider the exact same communication setup and repeat each experiment 3 times and plot the test accuracy curve over time until one of the algorithms stops. Figure 2 **(b)&(c)** compares the performance and speed of these algorithms.

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 $\delta > 0$,

$$\nabla^2 f_i(w) u \approx \frac{\nabla f_i(w + \delta u) - \nabla f_i(w - \delta u)}{\delta}. \tag{226}$$

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



(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 $n = 50$ 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, AFL, 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.