Journal of Machine Learning Research 24 (2023) 1-46 Submitted 11/21; Revised 12/22; Published 4/23

Memory-Based Optimization Methods for Model-Agnostic Meta-Learning and Personalized Federated Learning

Bokun Wang BOKUN-WANG@TAMU.EDU
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA

Zhuoning Yuan ZHUONING-YUAN@UIOWA.EDU
Department of Computer Science
The University of Iowa
Iowa City, IA 52242, USA

Yiming Ying YYING@ALBANY.EDU
Department of Mathematics and Statistics
University at Albany
Albany, NY 12222, USA

Tianbao Yang TIANBAO-YANG@TAMU.EDU
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA

Editor: Zaid Harchaoui

Abstract

In recent years, model-agnostic meta-learning (MAML) has become a popular research area. However, the stochastic optimization of MAML is still underdeveloped. Existing MAML algorithms rely on the "episode" idea by sampling a few tasks and data points to update the meta-model at each iteration. Nonetheless, these algorithms either fail to guarantee convergence with a constant mini-batch size or require processing a large number of tasks at every iteration, which is unsuitable for continual learning or cross-device federated learning where only a small number of tasks are available per iteration or per round. To address these issues, this paper proposes memory-based stochastic algorithms for MAML that converge with vanishing error. The proposed algorithms require sampling a constant number of tasks and data samples per iteration, making them suitable for the continual learning scenario. Moreover, we introduce a communication-efficient memory-based MAML algorithm for personalized federated learning in cross-device (with client sampling) and cross-silo (without client sampling) settings. Our theoretical analysis improves the optimization theory for MAML, and our empirical results corroborate our theoretical findings. Interested readers can access our code at `https://github.com/bokun-wang/moml`.

Keywords: Meta-Learning, Federated Learning, Model-Agnostic Meta-Learning, Personalized Federated Learning, Memory-Based Algorithms

1. Introduction

Despite the remarkable success of modern deep learning approaches, they are often criticized for their heavy reliance on large amounts of data (Marcus, 2018). In contrast, humans can learn with relatively small amounts of data thanks to their ability to continuously learn from multiple tasks. Recently, meta-learning has garnered significant attention for its ability to perform well on new tasks using the adaptation and prior knowledge gained from previous tasks (Schmidhuber, 1987; Thrun and Pratt, 2012; Hospedales et al., 2020). Among meta-learning approaches, the model-agnostic meta-learning (MAML) technique based on gradient-based optimization (Finn et al., 2017) has proven to be successful across a broad range of problems that can be trained using gradient descent. Specifically, MAML proposes to solve the following optimization problem.

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})), \tag{1}$$

where we use $\mathbf{w} \in \mathbb{R}^d$ to represent the meta-model, and n to denote the number of tasks. The risk function for the i-th task is denoted by $\mathcal{L}_i$, and can be expressed as $\mathcal{L}_i(\mathbf{w}) = \mathbb{E}_{\mathbf{z}\sim\mathbf{D}_i}[\ell_i(\mathbf{w}, \mathbf{z})]$. Here, $\mathbf{D}_i$ represents the data distribution for the i-th task, while $\ell_i(\cdot)$ denotes the loss function. The inner gradient step, $\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})$, represents an adaptation from the meta-model $\mathbf{w}$ to the i-th task.

MAML has received considerable attention from researchers, with several studies investigating its applications and extensions (Nichol et al., 2018; Antoniou et al., 2018; Behl et al., 2019; Yoon et al., 2018; Raghu et al., 2019; Li et al., 2017; Grant et al., 2018). However, the stochastic optimization algorithms used to solve the MAML problem in Eq. (1) are still far from satisfactory. Two key quantities are present in each "episode" of the optimization algorithms: the number of sampled tasks denoted by B, and the number of sampled data points per task denoted by K. The episode is designed to mimic the few-shot task by sub-sampling both tasks and data points (Vinyals et al., 2016; Snell et al., 2017; Ravi and Larochelle, 2017). Unfortunately, the original MAML method based on the episode does not necessarily converge to a stationary point of the objective function F in Eq. (1) unless K is sufficiently large. Recently, Fallah et al. (2020a) provided the first convergence analysis of the original MAML approach, which suggests that to find an ϵ-stationary point $\mathbf{w}$ satisfying $\mathbb{E}[\|\nabla F(\mathbf{w})\|] \leq \epsilon$, one needs to run MAML for $T = \mathcal{O}(1/\epsilon^2)$ iterations and sample $K = \mathcal{O}(1/\epsilon^2)$ data points for all n tasks in each iteration. However, these batch sizes are impractical for driving the error level ϵ to be sufficiently small.

In recent work, Hu et al. (2020) introduced two biased stochastic methods, namely BSGD and BSpiderBoost, which are modifications of the original MAML algorithm (Finn et al., 2017) for solving Eq. (1). These methods have convergence guarantees for finding an ϵ-stationary point, but require impractical settings for the number of sampled tasks and data points per iteration (K and B) as well as the number of iterations (T), as $K = \mathcal{O}(1/\epsilon^2)$, $B = \mathcal{O}(1)$, and $T = \mathcal{O}(1/\epsilon^4)$ for BSGD, and $K = \mathcal{O}(1/\epsilon^2)$, $B = \sqrt{n}$, and $T = \mathcal{O}(1/\epsilon^2)$ for BSpiderBoost. Neither setting is practical for driving the error level ϵ to be sufficiently small, not to mention the imposed additional assumptions (see Table 1 for a more thorough comparison). These findings suggest that the original MAML approach may not converge to an accurate solution if a small batch size K is used. Other studies have approached the

optimization of Eq. (1) as a two-level compositional function (Chen et al., 2020; Tutunov et al., 2020) or bilevel optimization problem (Franceschi et al., 2018; Ji et al., 2020b; Chen et al., 2021), but these methods either require K to be very large or involve passing through all n tasks at each iteration.

Federated Learning (FL) is a framework for distributed learning over a federation of mobile devices (Konecnỳ et al., 2016; McMahan et al., 2017; Kairouz et al., 2021; Wang et al., 2021). In FL, the local data of each device cannot be shared with other devices or the central server. There are two important settings of FL: *cross-silo* and *cross-device*. The cross-silo FL is typically located at data centers, where the number of clients is limited (say dozens) and the clients are available at each iteration. In contrast, the cross-device FL is deployed on a network of mobile devices, where the number of clients is much larger than that of cross-silo FL, but few clients are available in each iteration. Personalized FL has drawn attention due to the challenges and questions posed by data heterogeneity. Recently, the connection between meta-learning and personalized federated learning (FL) has been noticed, since both tasks in meta-learning and clients in federated learning are heterogeneous (Jiang et al., 2019). The convergence theory of the federated variant of MAML has been developed in Fallah et al. (2020b).

This paper aims to improve MAML optimization by addressing the following question:

Can we design efficient stochastic optimization algorithms for MAML, which can converge to a stationary point with only $K = \mathcal{O}(1)$ and $B = \mathcal{O}(1)$ to update the model?

1.1 Contributions

We present the main contributions of our work below:

- We address the problem of stochastic optimization for MAML in both single-node and federated learning settings. In the single-node setting, the server has access to all tasks and their data, while in the federated learning setting, the central server has no access to the individual tasks and their data on distributed clients. We propose two memory-based stochastic algorithms, MOML and LocalMOML. MOML is designed for the centralized setting, while LocalMOML can be used in both centralized and federated learning settings. The proposed algorithms maintain and update individualized models, or memories, for each task using a Momentum update. This involves computing a moving average of historical stochastic updates of individual models.

- We provide the convergence guarantees of MOML and LocalMOML for finding a stationary point of the non-convex objective with only $K = \mathcal{O}(1)$ data samples per task and $B = \mathcal{O}(1)$ tasks per iteration. To the best of our knowledge, this is the first work to achieve such results. We also provide a comparison of our theoretical results and key features of the proposed algorithms with other existing results in Table 1. Importantly, our LocalMOML algorithm consistently outperforms the existing Per-FedAvg algorithm (Fallah et al., 2020b) in terms of sample complexity.

- Our proposed methods MOML and LocalMOML support task/client sampling, which is a desirable property under both single-node learning and federated learning settings, unlike some methods listed in Table 1. Task sampling is desired when the tasks are on the same

Table 1: Comparison of proposed algorithms with existing approaches when the number of tasks n is finite. ϵ denotes the accuracy for an ϵ-stationary point $\mathbb{E}\left[\|\nabla F(\mathbf{w})\|\right] \leq \epsilon$. Ticks and crosses in **blue** are pros while those in **purple** are cons.

Single-Node Learning					
Algorithm	Task Sampling	Sample Complexity	#Data points (K) Per Iteration	Strict Assumptions	
				Bounded Gradient	Stochastic Lipschitz[(1)]
MAML (Fallah et al., 2020a)	✗	$\mathcal{O}\left(n\epsilon^{-4}\right)$	$\mathcal{O}(\epsilon^{-2})$	✗	✗
SCGD (Wang et al., 2017)	✗	$\mathcal{O}(n\epsilon^{-8})$	$\mathcal{O}(1)$	✓	✗
NASA (Ghadimi et al., 2020)	✗	$\mathcal{O}(n\epsilon^{-4})$	$\mathcal{O}(1)$	✓	✗
BSGD (Hu et al., 2020)	✓	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-2})$	✓[(2)]	✓
BSpiderBoost (Hu et al., 2020)	✓	$\mathcal{O}\left(n\epsilon^{-2}+\sqrt{n}\epsilon^{-4}\right)$	$\mathcal{O}(\epsilon^{-2})$	✓	✓
MOML$^{\mathbf{v1}}$ (This work)	✓	$\mathcal{O}\left(n\epsilon^{-5}\right)$	$\boldsymbol{\mathcal{O}(1)}$	✓	✗
MOML$^{\mathbf{v2}}$ (This work)	✓	$\mathcal{O}\left(n\epsilon^{-5}\right)$	$\boldsymbol{\mathcal{O}(1)}$	✗	✗
Personalized Federated Learning					
Algorithm	Client Sampling	Sample Complexity	Communication Complexity	Avg. #Data points (K) Per Iteration	
Per-FedAvg (Fallah et al., 2020b)	✗	$\mathcal{O}\left(n\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$	
Per-FedAvg **(This work)**[(3)]	✓	$\mathcal{O}(\epsilon^{-7})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	
LocalMOML (This work)	✗	$\boldsymbol{\mathcal{O}\left(n\epsilon^{-5}\right)}$	$\mathcal{O}(\epsilon^{-3})$	$\boldsymbol{\mathcal{O}(1)}$	
LocalMOML (This work)	✓	$\boldsymbol{\mathcal{O}\left(\epsilon^{-6}\right)}$	$\mathcal{O}(\epsilon^{-4})$	$\boldsymbol{\mathcal{O}(1)}$	

[(1)] Stochastic Lipschitz: $\nabla \ell(\cdot;\mathbf{z})$ is Lipschitz continuous for each $\mathbf{z}$, which is stronger than the Lipschitzness of $\mathcal{L}_i(\cdot)$ in Assumption 1.
[(2)] BSGD can obtain the same rate without the bounded gradient assumption if the weak convexity of $F(\cdot)$ is additionally assumed.
[(3)] The analysis of Per-FedAvg with client sampling in Fallah et al. (2020b) seems to be problematic. See Appendix E.3 for details.

Table 2: Comparison of proposed algorithms with existing approaches when the number of tasks n is infinite. For example, the tasks are online.

Single-Node Learning					
Algorithm	Sample Complexity	#Data points (K) Per Iteration	#Tasks (B) Per Iteration	Strict Assumptions	
				Bounded Gradient	Stochastic Lipschitz
MAML (Fallah et al., 2020a)	$\mathcal{O}\left(\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	✗	✗
BSpiderBoost (Hu et al., 2020)	$\mathcal{O}\left(\epsilon^{-5}\right)$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$ or $\mathcal{O}(\epsilon^{-1})$	✓	✓
BSGD (Hu et al., 2020)	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(1)$	✓	✓
LocalMOML (This work)	$\mathcal{O}(\epsilon^{-6})$	$\boldsymbol{\mathcal{O}(1)}$	$\boldsymbol{\mathcal{O}(1)}$	✓	✗

machine (single-node learning), as backpropagating through all n tasks requires much more GPU memory. Moreover, in the continual learning regime, only a small proportion of tasks might be available every iteration. In the cross-device federated learning regime, the server can only access available clients via the client sampling process because the number of clients is huge and direct connections cannot be easily established.

2. Related Works

In this section, we discuss previous works related to ours in four categories.

Meta-Agnostic Meta-Learning (MAML) Gradient-based MAML was introduced in Finn et al. (2017) and has since become a popular algorithm for learning from prior experience, with many applications in supervised learning, reinforcement learning, and more. Later on, several works have delved deeper into MAML to better understand its practical performance and provide some tricks of the trade for further improving its practicability (Antoniou et al., 2018; Raghu et al., 2019; Behl et al., 2019). The vanilla MAML has been generalized from various perspectives. For example, probabilistic MAML is introduced in Finn et al. (2018) to model a distribution over prior model parameters. Other algorithms with multi-step gradient descent (Ji et al., 2020c) and with partial parameters adaptation (Ji et al., 2020a) have also been proposed. Rajeswaran et al. (2019) proposed meta-learning with implicit gradients by formulating the problem as bilevel optimization. Besides, Hessian-free variants of MAML have been proposed to improve computational efficiency (Finn et al., 2017; Nichol et al., 2018; Zhou et al., 2019; Song et al., 2020; Fallah et al., 2020a).

Optimization theory of MAML In recent years, researchers have focused on addressing the computational and optimization challenges of MAML and its variants. For instance, Balcan et al. (2019) provided provable guarantees of a generalized framework of gradient-based MAML in the convex online learning scheme. When the loss function $\mathcal{L}_i$ is convex, global convergence of MAML has been established for meta-supervised learning and meta-reinforcement learning in Wang et al. (2020). When $\mathcal{L}_i$ is nonconvex, the convergence to stationary points of MAML and its first-order and Hessian-free variants is proved in Fallah et al. (2020a). While the BSGD algorithm proposed in Hu et al. (2020) has been shown to have theoretical and practical advantages over the results in (Fallah et al., 2020a), it relies on stronger assumptions. As previously noted, these results are still not entirely satisfactory for MAML. Finally, the convergence of iMAML to stationary points in the nonconvex setting has been demonstrated, but the theory requires processing all n tasks at each iteration. Besides, SCGD (Wang et al., 2017) and NASA (Ghadimi et al., 2020) can also be used to optimize the MAML objective since problem in Eq. (1) can be viewed as an instance of stochastic two-level compositional problems in the form of $\mathbb{E}_{\xi}[f_{\xi}(\mathbb{E}_{\xi'}[\mathbf{g}(\mathbf{w};\xi')];\xi)]$, where $\xi' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_N) \sim \mathbf{D}_1 \times \ldots \times \mathbf{D}_n$, $\xi = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \sim \mathbf{D}_1 \times \ldots \times \mathbf{D}_n$, $\mathbf{g}(\mathbf{w};\xi') = [\mathbf{w} - \alpha\nabla\ell_1(\mathbf{w};\mathbf{z}'_1); \cdots ; \mathbf{w} - \alpha\nabla\ell_n(\mathbf{w};\mathbf{z}'_n)] \in \mathbb{R}^{nd}$, and $f_{\xi}(\mathbf{g}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i([\mathbf{g}]_i, \mathbf{z}_i)$. However, a limitation of using SCGD or NASA to solve MAML is that they require passing through all n tasks at each iteration (Wang et al., 2017; Ghadimi et al., 2020). Additionally, these works assume that both ∇f and $\nabla \mathbf{g}$ are bounded.

Federated learning related to MAML Our work builds upon previous research that has explored the relationship between MAML and classical federated averaging (FedAvg) in federated learning (FL). In Jiang et al. (2019), the authors show how FedAvg can be connected to MAML and derive a heuristic-based algorithm that alternates between running FedAvg for several iterations and using a meta-learning approach for fine-tuning. This results in a good initial model for any client and improves personalized performance even when the local data is limited. Fallah et al. (2020b) show that the MAML-based Per-FedAvg leads to superior personalized federated learning performance compared to FedAvg on some

numerical experiments. Personalized federated learning similar to but not exactly the same as MAML has also been considered in several recent works (Hanzely and Richtárik, 2020; T. Dinh et al., 2020). Our work specifically focuses on federated learning with the vanilla MAML formulation, similar to Fallah et al. (2020b).

Continual learning Finally, it is worth noting that using a memory buffer to track each task has been explored in other continual learning paradigms to address the problem of catastrophic forgetting in a learning agent, such as memory-based lifelong learning (Lopez-Paz and Ranzato, 2017; Kirkpatrick et al., 2016; Guo et al., 2020). However, it is important to emphasize that the memory used in our MAML algorithms and that used in lifelong learning are distinct. In MAML, the memory is used to track individual models of different tasks, while in lifelong learning, it is used to store some training data for different tasks.

Finally, we note that the proposed techniques can be employed for solving other problems with similar structures to MAML, e.g., the meta-tailoring problem (Alet et al., 2021).

3. Preliminaries

In this section, we present the notation, assumptions, and key challenges in solving Eq. (1).

3.1 Notation

The Euclidean norm of a vector and the spectral norm of a matrix are denoted by $\|\cdot\|$. Calligraphic and capital letters, such as $\mathcal{B}$ and $\mathcal{S}$, denote sets. For a data distribution $\mathbf{D}$, we use $\mathcal{S} \sim \mathbf{D}$ to denote a set of i.i.d. samples following the distribution $\mathbf{D}$. We use $\mathbb{I}()$ to denote the indicator function. The unbiased stochastic gradient and stochastic Hessian of the risk function $\mathcal{L}_i$ based on a random set $\mathcal{S} \sim \mathbf{D}_i$ of size K are denoted by $\widehat{\nabla}_{\mathcal{S}}\mathcal{L}_i(\mathbf{w}) = \frac{1}{K}\sum_{\mathbf{z}_i \in \mathcal{S}} \nabla \ell_i(\mathbf{w}; \mathbf{z}_i)$, and $\widehat{\nabla}^2_{\mathcal{S}}\mathcal{L}_i(\mathbf{w}) = \frac{1}{K}\sum_{\mathbf{z}_i \in \mathcal{S}} \nabla^2 \ell_i(\mathbf{w}; \mathbf{z}_i)$, respectively. Refer to Table 6 for a complete list of notations used in this paper.

3.2 Assumptions

Throughout the paper, we assume that Assumption 1, 2, and 3 are satisfied, which are standard in the literature (Fallah et al., 2020a; Ji et al., 2020c; Rajeswaran et al., 2019).

Assumption 1 *$\mathcal{L}_i(\cdot)$ has L-Lipschitz continuous gradient and ρ-Lipschitz continuous Hessian, that is, $\|\nabla \mathcal{L}_i(\mathbf{w}) - \nabla \mathcal{L}_i(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$, and $\|\nabla^2 \mathcal{L}_i(\mathbf{w}) - \nabla^2 \mathcal{L}_i(\mathbf{w}')\| \leq \rho\|\mathbf{w} - \mathbf{w}'\|$ for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$.*

Assumption 2 *The variance of stochastic gradient $\nabla \ell(\mathbf{w}, \mathbf{z})$ and stochastic Hessian $\nabla^2 \ell(\mathbf{w}, \mathbf{z})$ are upper bounded:*

$$\mathbb{E}_{\mathbf{z}\sim\mathbf{D}_i}\left[\|\nabla \ell(\mathbf{w}; \mathbf{z}) - \nabla \mathcal{L}_i(\mathbf{w})\|^2\right] \leq \sigma_G^2, \quad \mathbb{E}_{\mathbf{z}\sim\mathbf{D}_i}\left[\left\|\nabla^2 \ell(\mathbf{w}; \mathbf{z}) - \nabla^2 \mathcal{L}_i(\mathbf{w})\right\|^2\right] \leq \sigma_H^2.$$

Assumption 3 *F is bounded below, $\inf_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) > -\infty$.*

Most of the existing results to solve Eq. (1) in the literature (Finn et al., 2017; Rajeswaran et al., 2019; Wang et al., 2017; Ghadimi et al., 2020; Chen et al., 2020; Fallah et al., 2020b) are under the bounded gradient assumption (Assumption 4).

Assumption 4 *There exists $G > 0$, $\|\nabla \mathcal{L}_i(\mathbf{w})\| \leq G$ for any $\mathbf{w} \in \mathbb{R}^d$.*

Instead of Assumption 4, Fallah et al. (2020a) establish the convergence theory of MAML based on Assumption 5, where the gradients are not necessarily bounded.

Assumption 5 *There exists $\gamma_G \geq 0$, $\frac{1}{n}\sum_{i=1}^n \|\nabla \mathcal{L}_i(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w})\|^2 \leq \gamma_G^2$ for all $\mathbf{w} \in \mathbb{R}^d$ and $\mathcal{L}(\mathbf{w}) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathcal{L}_i(\mathbf{w})$.*

3.3 Main Challenges

A key to the design of stochastic optimization of Eq. (1) is to estimate the gradient of the objective $\nabla F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}))\nabla \mathcal{L}_i(\mathbf{w} - \alpha \nabla \mathcal{L}_i(\mathbf{w}))$ based on random samples. Existing algorithms, such as those proposed in (Fallah et al., 2020a; Hu et al., 2020), typically estimate the gradient of the objective via mini-batch averaging:

$$\widehat{\Delta}_{\mathcal{B}} = \frac{1}{B}\sum_{i \in \mathcal{B}} (I - \alpha \widehat{\nabla}^2_{\mathcal{S}_2^i} \mathcal{L}_i(\mathbf{w}))\widehat{\nabla}_{\mathcal{S}_3^i} \mathcal{L}_i(\mathbf{w} - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w})), \tag{2}$$

where $\mathcal{B}$ denotes the set of B sampled tasks, $\mathcal{S}_1^i, \mathcal{S}_2^i, \mathcal{S}_3^i$ denote three independent sample sets of size K for each sampled task $\mathcal{T}_i$. However, this naïve approach could lead to a large optimization error when K is not large enough.

Thus, the first challenge is to design an algorithm that provably converges with sub-sampled tasks and a constant number of data points. To tackle this challenge, we borrow the idea from Wang et al. (2017) that keeps track of the sequence $\mathbf{v}_i(\mathbf{w}_t) = \mathbf{w}_t - \alpha \nabla \mathcal{L}_i(\mathbf{w}_t)$ with an estimator $\mathbf{u}^i$ for each task $\mathcal{T}_i$ (also known as the personalized model). The first novelty of our work, compared to prior work (Wang et al., 2017), lies in the task sampling approach, where the algorithm only needs to sample data and compute the stochastic gradients for a subset of B tasks, instead of all n tasks.

Moreover, it is even more challenging to establish similar convergence guarantees without the bounded gradient assumption (Assumption 4). As shown in Lemma 10, the gradient-Lipschitz parameter of the meta-objective $F(\mathbf{w})$ is $L(\mathbf{w}) \coloneqq 4L + \frac{2\rho\alpha}{n}\sum_{i=1}^n \|\nabla \mathcal{L}_i(\mathbf{w})\|$. To handle the unbounded gradients, Fallah et al. (2020a) estimate the gradient-Lipschitz parameter $L(\mathbf{w}_t)$ by a stochastic estimator $\widehat{L}(\mathbf{w}_t) \coloneqq 4L + \frac{2\rho\alpha}{|\mathcal{B}_{L_t}|}\sum_{i \in \mathcal{B}_{L_t}} \left\|\widehat{\nabla}_{\mathcal{S}_{L_t}^i} \mathcal{L}_i(\mathbf{w}_t)\right\|$ and set the stepsize η_t to be inversely proportional to $\widehat{L}(\mathbf{w}_t)$. However, MAML with that stepsize η_t still requires $K = \mathcal{O}(1/\epsilon^2)$ data samples per task in each iteration to ensure convergence. Thus, it was still an open problem whether the proposed technique can be extended to this setting for getting rid of the unrealistic requirement of large batch size.

4. Memory-Based MAML (MOML) in the Single-Node Learning

We tackle the challenges mentioned in Section 3.3 by proposing the MOML algorithm.

Algorithm 1 MOML$^{\text{v1}}$

Hyperparameters: β (suggested value 0.5), η (to be tuned in practice)
for $t = 0, 1, \ldots, T-1$ **do**
  Select a batch of B tasks $\mathcal{B}_t$ from n tasks
  for each task $\mathcal{T}_i$, $i \in \mathcal{B}_t$ **do**
    Select K samples $\mathcal{S}_1^i \sim \mathbf{D}_i$ and compute $\widehat{\mathbf{v}}_t^i = \mathbf{w}_t - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w}_t)$
    Update the personalized model by $\mathbf{u}_{t+1}^i = (1-\beta)\mathbf{u}_t^i + \beta \widehat{\mathbf{v}}_t^i$.
  end for
  Select K samples $\mathcal{S}_2^i$ and $\mathcal{S}_3^i$ from $\mathbf{D}_i$ of task $\mathcal{T}_i$, $i \in \mathcal{B}_t$ and compute $\widehat{\Delta}_{\mathcal{B}_t}$ by (3)
  Update the meta-model by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \widehat{\Delta}_{\mathcal{B}_t}$
end for

4.1 Algorithm Outline

The proposed MOML$^{\text{v1}}$ (Algorithm 1) updates the personalized model of a sampled task $\mathcal{T}_i$, $i \in \mathcal{B}_t$ by a momentum step while those of the other tasks $i \notin \mathcal{B}_t$ are untouched, that is,

$$\mathbf{u}_{t+1}^i = \begin{cases} (1-\beta_t)\mathbf{u}_t^i + \beta_t \widehat{\mathbf{v}}_t^i & i \in \mathcal{B}_t \\ \mathbf{u}_t^i & i \notin \mathcal{B}_t \end{cases}, \tag{v1}$$

where $\beta \in (0, 1]$ is the momentum factor. It is worth noting that MOML$^{\text{v1}}$ with $\beta = 1$ recovers the original MAML algorithm. Based on the updated personalized models $\mathbf{u}_{t+1}^i$, we can compute the stochastic gradient by

$$\widehat{\Delta}_{\mathcal{B}_t} = \frac{1}{B} \sum_{i \in \mathcal{B}_t} (I - \alpha \widehat{\nabla}^2_{\mathcal{S}_2^i} \mathcal{L}_i(\mathbf{w}_t)) \widehat{\nabla}_{\mathcal{S}_3^i} \mathcal{L}_i(\mathbf{u}_{t+1}^i). \tag{3}$$

What is the intuition behind (v1) and (3)? When the batch size $|\mathcal{S}_1^i|$ is not large enough, the estimator $\widehat{\mathbf{v}}_t^i \coloneqq \mathbf{w} - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w})$ to compute $\widehat{\Delta}_{\mathcal{B}}$ in (2) might lead to large error to estimate $\mathbf{w} - \alpha \nabla \mathcal{L}_i(\mathbf{w})$ and impede the convergence. Instead, we design a new personalized model estimator $\mathbf{u}_t^i$ which is an exponential moving average of many "historical estimators" $\widehat{\mathbf{v}}_{t'}^i$ in the past iterations $t' < t$, which covers much more data points and its estimation error is provably small. Please refer to the discussion after Lemma 2 for a formal justification.

4.2 Convergence Analysis

We establish the convergence guarantees of MOML$^{\text{v1}}$ based on Assumption 1, 2, 3 and 4. With these assumptions, the meta-objective $F(\cdot)$ is L_F-smooth (see Lemma 10). Based on this fact, we can derive the lemma below.

Lemma 1 *If $\alpha \in (0, 1/L]$, the iterates $\{\mathbf{w}_t\}_{t=0}^{T-1}$ of* MOML$^{\text{v1}}$ *satisfy that*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2 \right] \le \frac{2F(\mathbf{w}_0)}{\eta T} + \frac{\eta L_F}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \widehat{\Delta}_{\mathcal{B}_t} \right\|^2 \right] + \frac{8L^2}{BT} \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{i \in \mathcal{B}_t} \left\| \mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i \right\|^2 \right].$$

Next, we need to show the error of tracking $\mathbf{v}_i(\mathbf{w}_t)$ (the last term in Lemma 1) is vanishing when $\beta \in (0,1)$. Different from existing analysis of stochastic compositional optimization (Wang et al., 2017), the estimators $\mathbf{u}^i$ for tracking the inner functions $\mathbf{v}(\mathbf{w}) \coloneqq ([\mathbf{v}_i(\mathbf{w})]^\top, \ldots, [\mathbf{v}_n(\mathbf{w})]^\top)^\top$ are only partially updated due to the task sampling. Hence, we need a different technique to bound the error. Given the total number of iterations T, we define n totally ordered sets $\mathbb{T}_1, \ldots, \mathbb{T}_n$, where $\mathbb{T}_i \subseteq [T]$ contains the iteration indices that the i-th task is sampled, that is to say, $\mathbb{T}_i = \{t_1^i, \ldots, t_k^i, \ldots, \}$. If task $\mathcal{T}_i$ is sampled in iteration t and the index of t in $\mathbb{T}_i$ is k, then $t_k^i = t$. Hence, we can define a mapping from t to k for $i \in \mathcal{B}_t$. Based on this definition, we can obtain that

$$\sum_{t=0}^{T-1} \sum_{i \in \mathcal{B}_t} \left\| \mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i \right\|^2 = \sum_{i=1}^{n} \sum_{k=0}^{T_i - 1} \left\| \mathbf{v}_i(\mathbf{w}_{t_k^i}) - \mathbf{u}_{t_k^i+1}^i \right\|^2. \tag{4}$$

Here T_i is the cardinality of set $\mathbb{T}_i$, which is random and depends on the task sampling $\{\mathcal{B}_t\}_{t=0}^{T-1}$. Lemma 2 upper bounds the right-hand side of (4).

Lemma 2 *Suppose that the batch of tasks $\mathcal{B}_t$ is sampled uniformly at random. The error of* MOML$^{\text{v1}}$ *with $|\mathcal{S}_1^i| = K$ to keep track of $\mathbf{v}_i(\mathbf{w}_t)$ can be upper bounded as*

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \sum_{i \in \mathcal{B}_t} \left\| \mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i \right\|^2 \right] \leq \left(\frac{n\sigma_G^2}{\beta K T} + \frac{16\eta^2 n^2 C_\Delta}{\beta^2 B^2} \right) \mathbb{I}[\beta \in (0,1)] + \frac{\beta \alpha^2 \sigma_G^2}{K}, \tag{5}$$

where $C_\Delta \coloneqq (\alpha^2 \sigma_H^2 / K + (1+\alpha L)^2)(\sigma_G^2/K + G^2)$.

Lemmas 1 and 2 explain why our MOML algorithm converges with $B = \mathcal{O}(1)$ tasks and $K = \mathcal{O}(1)$ samples while previous works on MAML do not. As shown in Lemma 2, MOML's estimation error $\mathbb{E}\left[\frac{1}{B} \sum_{i \in \mathcal{B}_t} \left\| \mathbf{v}_i(\mathbf{w}_t) - \widehat{\mathbf{v}}_t^i \right\|^2 \right] \leq \mathcal{O}(\epsilon^2)$ even with $K = \mathcal{O}(1)$ by setting $\eta = \mathcal{O}(\epsilon^3)$, $\beta = \mathcal{O}(\epsilon^2)$. For the convergence of MAML, both a) the product of stepsize η and the second moment of the stochastic meta-gradient $\hat{\Delta}_{\mathcal{B}_t}$ and b) the estimation error of personal models $\widehat{\mathbf{v}}_t^i = \mathbf{w}_t - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w}_t)$ should be small. To be specific, it needs

$$\eta \mathbb{E}\left[\left\| \hat{\Delta}_{\mathcal{B}_t} \right\|^2 \right] \leq \mathcal{O}(\epsilon^2) \quad \text{and} \quad \mathbb{E}\left[\frac{1}{B} \sum_{i \in \mathcal{B}_t} \left\| \mathbf{v}_i(\mathbf{w}_t) - \widehat{\mathbf{v}}_t^i \right\|^2 \right] \leq \mathcal{O}(\epsilon^2) \tag{$\star$}$$

for an ϵ-stationary point. Fallah et al. (2020a) on MAML sets $\eta = \mathcal{O}(1)$ and bounds the second moment as follows.

$$\mathbb{E}\left[\left\| \widehat{\Delta}_{\mathcal{B}_t} \right\|^2 \right] \leq \begin{cases} \underbrace{2\left(1 + \frac{20}{B}\right) \mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2 \right]}_{\coloneqq \ddagger} + \frac{\sigma_G^2}{K} + \frac{1}{B}\left(14G^2 + \frac{3\sigma_G^2}{K}\right) & \text{total \#tasks n is infinite} \\ \underbrace{2\left(1 + \frac{20}{B}\right) \mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2 \right]}_{\coloneqq \ddagger} + \frac{\sigma_G^2}{K} + \frac{(n-B)}{B(n-1)}\left(14G^2 + \frac{3\sigma_G^2}{K}\right) & \text{total \#tasks n is finite.} \end{cases}$$

The ‡ term can be canceled out with the L.H.S. of Lemma 1. Besides, MAML (i.e. MOML with $\beta = 1$) also satisfies

$$\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left\|\mathbf{v}_i(\mathbf{w}_t) - \widehat{\mathbf{v}}_t^i\right\|^2\right] \leq \frac{\alpha^2\sigma_G^2}{K}. \tag{$\diamond$}$$

To make ($\star$) hold, Fallah et al. (2020a) need $B = \mathcal{O}(\epsilon^{-2})$, $K = \mathcal{O}(\epsilon^{-2})$ for infinite n case while $B = n$, $K = \mathcal{O}(\epsilon^{-2})$ for finite n case. More recent work Hu et al. (2020) instead use a small stepsize $\eta = \mathcal{O}(\epsilon^2)$ for MAML. Thus, making ($\star$) hold only needs $B = \mathcal{O}(1)$ tasks. However, it still needs $K = \mathcal{O}(\epsilon^{-2})$ data points due to ($\star$) and ($\diamond$). It is not clear how to remove the $K = \mathcal{O}(\epsilon^{-2})$ requirement for vanilla MAML. Next, we are ready to present the main convergence theorem of MOML$^{\text{v1}}$.

Theorem 3 (Informal) *Under Assumptions 1, 2 3, 4,* MOML$^{\text{v1}}$ *with stepsizes* $\eta_t = \eta = \mathcal{O}(\epsilon^{-3})$, $\beta_t = \beta = \mathcal{O}(\epsilon^{-2})$ *and constant batch sizes* $|\mathcal{S}_1^i| = |\mathcal{S}_2^i| = |\mathcal{S}_3^i| = K = \mathcal{O}(1)$, $|\mathcal{B}_t| = B = \mathcal{O}(1)$ *can find a stationary point* $\mathbf{w}_\tau$ *in* $T = \mathcal{O}(n\epsilon^{-5})$ *iterations.*

Compared to previous works that aim to solve Eq. (1) under the same assumptions, MOML$^{\text{v1}}$ can ensure convergence without the need for an extremely large batch $K = \mathcal{O}(1/\epsilon^2)$, as required in (Hu et al., 2020), or processing all n tasks (Ghadimi et al., 2020).

4.3 Handling Unbounded Gradients

We propose another variant of MOML — MOML$^{\text{v2}}$ in Algorithm 2 that is provably convergent with possibly unbounded gradients and only requires $K = \mathcal{O}(1)$ data samples for each sampled task in one iteration. In MOML$^{\text{v2}}$, the personalized model is updated as $\mathbf{u}^i$

$$\mathbf{u}_{t+1}^i = \begin{cases}(1-\beta_t)\mathbf{u}_t^i + \beta_t\mathbf{w}_t + \frac{\beta_t}{p_i}(\widehat{\mathbf{v}}_t^i - \mathbf{w}_t) & i \in \mathcal{B}_t' \\ (1-\beta_t)\mathbf{u}_t^i + \beta_t\mathbf{w}_t & i \notin \mathcal{B}_t'\end{cases}, \tag{v2}$$

where $\mathcal{B}_t'$ is independent of $\mathcal{B}_t$ and p_i is the probability of selecting task i, that is, $p_i = \text{Prob}(i \in \mathcal{B}_t')$. Besides, we set $\eta_t = \frac{\eta_0}{\widehat{L}(\mathbf{w}_t)}$ and $\beta_t = 6L^2\eta_0^{-1/3}\eta_{t-1}$ for MOML$^{\text{v2}}$. Under the same set of assumptions as Fallah et al. (2020a), we can show the error of tracking the inner function $\mathbf{v}_i(\mathbf{w}_t)$ is diminishing for MOML$^{\text{v2}}$ when η_0 is properly chosen.

Lemma 4 *If* $\eta_0 \leq \min\{(1/3L)^{3/2}, (3L^2/C_3C_6)^{3/2}\}$, *we have*

$$\mathbb{E}\left[\Upsilon_{t+1} \mid \mathcal{F}_t\right] \leq \left(1 - 3L^2\eta_0^{-\frac{1}{3}}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]\right)\mathbb{E}\left[\Upsilon_t \mid \mathcal{F}_t\right] + \eta_0^{\frac{1}{3}}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]C_9\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_0^{\frac{4}{3}}C_{10},$$

where $\Upsilon_t \coloneqq \frac{1}{n}\sum_{i=1}^n\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2$ *and* C_9, C_{10} *are* $\mathcal{O}(1)$ *constants w.r.t.* ϵ *and* n [1].

Theorem 5 (Informal) *Under Assumptions 1, 2, 3, and 5, it is guaranteed that* MOML$^{\text{v2}}$ *with* $\eta_t = \frac{\eta_0}{\widehat{L}(\mathbf{w}_t)}$, $\beta_t = 6L^2\eta_0^{-1/3}\eta_{t-1}$, $\eta_0 = \mathcal{O}(\epsilon^{-3})$ *and constant batch sizes* $|\mathcal{S}_1^i| = |\mathcal{S}_2^i| = |\mathcal{S}_3^i| = K = \mathcal{O}(1)$, $|\mathcal{B}_t| = |\mathcal{B}_t'| = B = \mathcal{O}(1)$ *can find an* ϵ*-stationary point* $\mathbf{w}_\tau$ *in* $T = O(C_p\epsilon^{-5})$ *iterations, where* $C_p = \max_i 1/p_i - 1$.

Algorithm 2 MOML$^{\text{v2}}$

Hyperparameters: β (suggested value 0.5), η (to be tuned in practice)
for $t = 0, 1, \ldots, T-1$ **do**
  Select two mutually independent batches of tasks $\mathcal{B}_t$, $\mathcal{B}'_t$ from n tasks
  for each task $\mathcal{T}_i$, $i \in \mathcal{B}'_t$ **do**
    Select K samples $\mathcal{S}_1^i \sim \mathbf{D}_i$ and compute $\widehat{\mathbf{v}}_t^i = \mathbf{w}_t - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w}_t)$
  end for
  Update the personalized model $\mathbf{u}^i$ by

$$\mathbf{u}_{t+1}^i = \begin{cases} (1-\beta_t)\mathbf{u}_t^i + \beta_t \mathbf{w}_t + \frac{\beta_t}{p_i}(\widehat{\mathbf{v}}_t^i - \mathbf{w}_t) & i \in \mathcal{B}'_t \\ (1-\beta_t)\mathbf{u}_t^i + \beta_t \mathbf{w}_t & i \notin \mathcal{B}'_t \end{cases}.$$

  Select K samples $\mathcal{S}_2^i$ and $\mathcal{S}_3^i$ from $\mathbf{D}_i$ of task $\mathcal{T}_i$, $i \in \mathcal{B}_t$ and compute $\widehat{\Delta}_{\mathcal{B}_t}$ by (3)
  Update the meta-model by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \widehat{\Delta}_{\mathcal{B}_t}$
end for

Remark 6 *Theorem 5 demonstrates that* MOML$^{\text{v2}}$ *improves upon the theory of MAML (Fallah et al., 2020a) by eliminating the need to sample $K = \mathcal{O}(1/\epsilon^2)$ data samples in each iteration. However, it should be noted that both the* MAML *variant in Fallah et al. (2020a) and* MOML$^{\text{v2}}$ *are more of theoretical interest because they require additional tasks/samples $\mathcal{B}_{\mathcal{L}_t}$ and $\mathcal{S}_{L_t}$ to estimate the gradient-Lipschitz parameter and then calculate the step size η_t in each iteration. Previous research (Ji et al., 2020c) and our experiments have demonstrated that the gradients $\nabla \mathcal{L}_i(\mathbf{w}_t)$ are well-bounded during the meta-training process, making* MOML$^{\text{v1}}$ *with a constant step size a more practical variant.*

5. LocalMOML for Personalized Federated Learning

This section presents a stochastic algorithm for solving MAML in the federated learning setting, assuming that there are n clients, and the i-th task and its corresponding data are only accessible at the i-th client. Note that this assumption can be relaxed to a setting where the n tasks are allocated to $m < n$ clients with non-overlapping. These clients are only permitted to aggregate the models and not exchange any data. However, a naive implementation of MOML in the federated learning setting would require aggregating the local gradient estimator $\widehat{\Delta}_{\mathcal{B}}^i$ at every iteration, or equivalently, aggregating the local copies of the meta-model at every iteration. Consequently, the communication complexity would be as high as the iteration complexity $T = \mathcal{O}(n/\epsilon^5)$. Our objective is to reduce communication complexity by proposing communication-efficient federated learning algorithms.

5.1 Algorithm Outline

Our algorithm, called LocalMOML, is presented in Algorithm 3. This algorithm is partially motivated by the numerous algorithms in federated learning that use local computations to trade-off communications (Yang, 2013; McMahan et al., 2017; Deng et al., 2020; Karimireddy

1. Proof of Lemma 4 in the appendix specifies the detailed expressions of C_9 and C_{10}.

Algorithm 3 LocalMOML

Hyperparameters: β (suggested value 0.5), η (to be tuned in practice)
for $r = 1, \ldots, R$ **do**
  Select a batch of B clients $\mathcal{B}_r$.
  for each client $\mathcal{C}_i$, $i \in \mathcal{B}_r$ **do**
    if client sampling **then**
      Select K_0 samples $\mathcal{S}_0^i \sim \mathbf{D}_i$, reset $\mathbf{u}_{r,1}^i = \mathbf{w}_{r,1}^i - \alpha \widehat{\nabla}_{\mathcal{S}_0^i} \mathcal{L}_i(\mathbf{w}_{r,1}^i)$
    else
      Set $\mathbf{u}_{r,1}^i = \mathbf{u}_{r-1,H}^i$
    end if
    for $h = 1, \ldots, H$ **do**
      Sample $\mathcal{S}_1^i$ to update the personalized model $\mathbf{u}_{r,h}^i$ by (6)
      Select two sets $\mathcal{S}_2^i$ and $\mathcal{S}_3^i$ of from $\mathbf{D}_i$ to compute $\widehat{\Delta}_{r,h}^i$ by (7), and update the local model by $\mathbf{w}_{r,h+1}^i = \mathbf{w}_{r,h}^i - \eta \widehat{\Delta}_{r,h}^i$
    end for
    Client $\mathcal{C}_i$ sends $\mathbf{w}_{r,H+1}^i$ to the server.
  end for
  The server aggregates and broadcasts $\mathbf{w}_{r+1} = \frac{1}{B} \sum_{i \in \mathcal{B}_r} \mathbf{w}_{r,H+1}^i$
end for

et al., 2019; Stich, 2019; Lin et al., 2020; Woodworth et al., 2020; Khaled et al., 2020). Each client not only maintains and updates its personalized model $\mathbf{u}^i$ but also maintains and updates its local copy of the meta-model denoted by $\mathbf{w}^i$. A key feature of LocalMOML is that H local steps are run on each sampled client before these clients communicate to aggregate the local meta-models.

We consider both the cross-silo setting and the cross-device setting in the literature on federated learning. In the cross-device setting only a partial set of $B < n$ clients are sampled to update their local models at each round, while in the cross-silo setting all n clients will participate in updating the model at each round, that is, $B = n$. In these two settings, LocalMOML needs different ways to initialize the personalized model $\mathbf{u}_{r,h}^i$ at the beginning of each round. In the cross-silo setting ($B = n$), personalized models are directly copied from the end of the previous round, that is, $\mathbf{u}_{r,1}^i = \mathbf{u}_{r-1,H}^i$. In the cross-device setting, ($B < n$), the personalized models for the sampled tasks are restarted as $\mathbf{u}_{r,1}^i = \mathbf{w}_{r,1}^i - \alpha \widehat{\nabla}_{\mathcal{S}_0^i} \mathcal{L}_i(\mathbf{w}_{r,1}^i)$ for $i \in \mathcal{B}_r$. Then, the personalized model $\mathbf{u}_{r,h}^i$ for a sampled client i is updated as:

$$\mathbf{u}_{r,h}^i = (1-\beta)\mathbf{u}_{r,h-1}^i + \beta \left(\mathbf{w}_{r,h}^i - \alpha \widehat{\nabla}_{\mathcal{S}_1^i} \mathcal{L}_i(\mathbf{w}_{r,h}^i) \right). \tag{6}$$

Based on the local personalized model $\mathbf{u}^i$ and meta model $\mathbf{w}^i$, the stochastic gradient estimator is computed as:

$$\widehat{\Delta}_{r,h}^i = (I - \alpha \widehat{\nabla}_{\mathcal{S}_2^i}^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i)) \widehat{\nabla}_{\mathcal{S}_3^i} \mathcal{L}_i(\mathbf{u}_{r,h}^i). \tag{7}$$

Once all H iterations have been completed in each round, the local copies of the meta-model at each client are aggregated for synchronization. We note that the Per-FedAvg

algorithm (Fallah et al., 2020b) is a special case of LocalMOML (Algorithm 3) with $\beta = 1$, and that the original MAML algorithm corresponds to LocalMOML with $\beta = 1$ and $H = 1$.

It is worth noting that LocalMOML can also be implemented in the single-node learning setting, for example by parallelizing on multiple CPU cores of a single machine. While MOML needs to maintain individualized models for all n tasks during the entire training process, which limits its scalability to a large number of tasks, LocalMOML only needs to maintain individualized models for a small subset of B sampled tasks within each round (epoch) r, where B can be small enough to avoid critical memory issues. After completing the iterations of round r, the individualized model $\mathbf{u}_{r,H}^i$ for the sampled task in round r can be erased from memory and will not be used in any later round $r' > r$, even if that task is selected again. This means that LocalMOML can handle an infinite number of tasks without encountering memory constraints.

5.2 Convergence results of LocalMOML

We analyze LocalMOML under the same assumptions as the Per-FedAvg (Fallah et al., 2020b), that is, Assumptions 1, 2, 3, 4, 5, and the assumption below.

Assumption 6 *There exists* $\gamma_H \geq 0$ *that:* $\frac{1}{n}\sum_{i=1}^n \left\|\nabla^2 \mathcal{L}_i(\mathbf{w}) - \frac{1}{n}\sum_{i=1}^n \nabla^2 \mathcal{L}_i(\mathbf{w})\right\|^2 \leq \gamma_H^2$.

Note that Assumptions 1 and 4 implies Assumptions 5 and 6. However, directly utilizing Assumptions 5 and 6 in the analysis can explicitly show the impact of the dissimilarity of data distribution on the final performance (Fallah et al., 2020b).

Theorem 7 *R rounds of* LocalMOML *with the stepsize* $\eta \leq \min\left\{\frac{C_4}{H}, \frac{C_5\beta}{\mathbb{I}[\beta\in(0,1)]}\right\}$ *leads to*

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^R \mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right] \\
&\leq \frac{4F(\mathbf{w}_1)}{\eta T} + 32\eta^2 H(H-1)C_1(\hat{\sigma}^2 + 2\gamma_F^2) + \frac{4\eta}{B}\left(\hat{\sigma}^2 + \frac{4(n-B)}{(n-1)}H\gamma_F^2\right) + \frac{64C_3\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&+ \frac{64C_3\sigma_G^2}{\beta\left(\mathbb{I}[B<n]HS_0 + \mathbb{I}[B=n]HR\right)}\mathbb{I}[\beta\in(0,1)] + \frac{3072C_3\eta^2(\hat{\sigma}^2+2\gamma_F^2)}{\beta^2}\mathbb{I}[\beta\in(0,1)],
\end{aligned}$$

where $T = RH$, $\hat{\sigma}^2 \coloneqq \frac{2\sigma_G^2}{|\mathcal{S}_3^i|} + \frac{2\alpha^2\sigma_G^2}{|\mathcal{S}_3^i|}\left(\frac{\sigma_H^2}{|\mathcal{S}_2^i|} + L^2\right) + \frac{\alpha^2 G^2\sigma_H^2}{|\mathcal{S}_2^i|}$, *and* C_1, C_3, C_4, C_5 *are* $\mathcal{O}(1)$ *constants w.r.t.* ϵ *and* n.

We consider the special case $\alpha = 0$, $H = 1$, $\beta = 1$, $B = n$ where LocalMOML becomes SGD and Theorem 7 recovers the standard result of SGD: $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{F(\mathbf{w}_1)}{\eta T} + \eta\sigma_G^2\right)$. Besides, the term $\gamma_F^2 \coloneqq 3G^2\alpha^2\gamma_H^2 + 192\gamma_G^2$ in Theorem 7 explicitly shows the impact of the dissimilarity of data distribution on the final performance. To better understand the complexities of LocalMOML, we present two corollaries below corresponding to the cross-silo $B = n$ and cross-device $B = \mathcal{O}(1)$ settings.

Corollary 1 (Cross-device Setting) *In this setting, we set* $\beta = O(\epsilon^2), \eta = \mathcal{O}(\epsilon^4), H = \mathcal{O}(1/\epsilon^2)$, $|\mathcal{S}_0^i| = K_0, |\mathcal{S}_1^i| = |\mathcal{S}_2^i| = |\mathcal{S}_3^i| = K$, $K = O(1), K_0 = H$. *Then, we can conclude*

that $T = \mathcal{O}(1/\epsilon^6)$ *and* $R = \mathcal{O}(1/\epsilon^4)$*. Hence, the average number of data points per-iteration is* $(K_0 + 3HK)/H = \mathcal{O}(1)$*, the total sample complexity is* $KBT + RK_0 = \mathcal{O}(1/\epsilon^6)$ *and the communication complexity is* $R = \mathcal{O}(1/\epsilon^4)$*.*

Corollary 2 (Cross-silo Setting) *In this setting, we set* $\beta = \mathcal{O}(\epsilon^2), \eta = \mathcal{O}(\epsilon^3), H = \mathcal{O}(1/\epsilon^2)$*,* $|\mathcal{S}_0^i| = 0, |\mathcal{S}_1^i| = |\mathcal{S}_2^i| = |\mathcal{S}_3^i| = K$*, and* $K = \mathcal{O}(1)$*. Then, we can conclude that* $T = \mathcal{O}(1/\epsilon^5)$ *and* $R = \mathcal{O}(1/\epsilon^3)$*. Hence, the total sample complexity is* $KNT = \mathcal{O}(n/\epsilon^5)$ *and the communication complexity is* $R = \mathcal{O}(1/\epsilon^3)$*.*

Thus, in both cross-silo and cross-device settings, our LocalMOML not only removes the unrealistic requirement of the large batch size $\mathcal{O}(1/\epsilon^2)$ in every iteration for the convergence of Per-FedAvg but also improves the sample complexity.

6. Experiments

We evaluate the performance of our proposed algorithms, MOML$^{\text{v1}}$, MOML$^{\text{v2}}$ and LocalMOML, on sinewave regression and one-shot classification tasks in the single-node setting. Furthermore, we demonstrate the effectiveness of LocalMOML in the simulated federated learning setting for the image classification task.

6.1 Sinewave Regression in the Single-Node Setting

First, we compare our proposed MOML$^{\text{v1}}$, MOML$^{\text{v2}}$ and LocalMOML with baselines NASA, MAML (BSGD), BSpiderBoost, and Reptile (Nichol et al., 2018) on the sinewave regression problem (Finn et al., 2017) in the single-node setting. We generate 25 tasks for training in total, each of which is to fit the function $f(x) = A\sin(\phi + x)$, where $A = \{1, 2, 3, 4, 5\}$, $\phi = \frac{i\pi}{5}$, $i = 1, 2, 3, 4, 5$. Similar to Finn et al. (2017), we choose the feedforward neural network with ReLU nonlinearities and 2 hidden layers of size 40 as the model and the mean-square error as the loss function. The training and validation data of each task are randomly sampled in an online manner $x \sim \text{Unif}(-5, 5)$. NASA uses all n tasks while the others sample $B = 3$ out of n tasks. We consider two possible minibatch sizes of data points: $K = 1$ and $K = 3$ in one iteration. The meta-learned model is adapted to 5 randomly sampled unseen tasks $A \sim \text{Unif}(1, 5)$, $\phi \sim \text{Unif}(\frac{\pi}{5}, \pi)$, $x \in [-5, 5]$ after 10 steps of gradient descent with learning rate 0.01 and 10 test data points (in other words, 10-shots). After the adaptation, we evaluate the final test error on another 100 data points from the unseen task. The inner step size α is set to 0.01 for all algorithms. The outer step size η is decayed 10 times at 75% of the total iterations[2] and its initial value is tuned for the algorithms separately by grid search in $\{0.1, 0.05, 0.01, 0.005, 0.001\}$. We also tune β for MOML$^{\text{v1}}$, MOML$^{\text{v2}}$ and LocalMOML. It turns out that $\beta = 0.3$ and $\beta = 0.5$ work reasonably well for MOML$^{\text{v1}}$ and LocalMOML while $\beta = 0.1$ is good for MOML$^{\text{v2}}$. For LocalMOML, we set the size of the initial number of samples K_0 of each round to be 2 times K and $H = 5$. The results are averaged over 5 trials with different random seeds.

We compare the convergence of our proposed algorithms and the baselines in terms of the number of samples. The training error Eq. (1) is approximated by $100n$ data points sampled from n tasks. As seen in Figure 1, MOML$^{\text{v1}}$ converges the fastest among the

2. except for BSpiderBoost, which requires a $\mathcal{O}(1)$ step size in its theory.

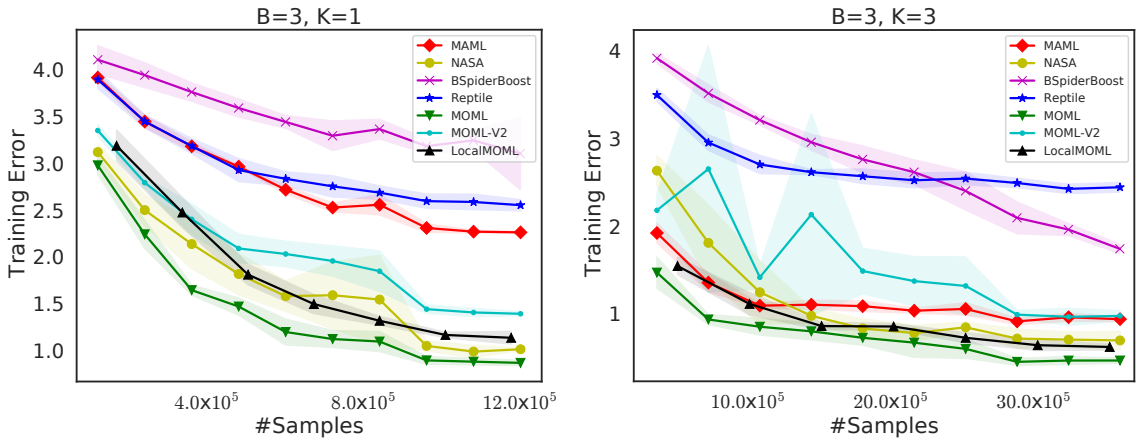


Figure 1: Convergence comparison in terms of the number of samples.

Table 3: Comparison of final test error for the sinewave regression task in the single-node setting. For each metric, the best one is highlighted in black and the second-best one is highlighted in gray.

$K = 1$							
Metrics	MAML (BSGD)	Reptile	NASA	BSpiderBoost	MOML^{v1}	MOML^{v2}	LocalMOML
Avg. Test Error	0.889 ± 0.021	1.212 ± 0.054	0.361 ± 0.010	1.300 ± 0.104	**0.291 ± 0.012**	0.448 ± 0.012	0.462 ± 0.043
Avg. Time Per Iteration (ms)	1.713 ± 0.030	**1.380 ± 0.071**	17.524 ± 1.306	3.960 ± 0.090	2.180 ± 0.031	7.863 ± 0.302	2.556 ± 0.100
$K = 3$							
Metrics	MAML (BSGD)	Reptile	NASA	BSpiderBoost	MOML^{v1}	MOML^{v2}	LocalMOML
Avg. Test Error	0.321 ± 0.015	1.132 ± 0.028	0.214 ± 0.050	0.650 ± 0.042	0.196 ± 0.068	0.268 ± 0.026	**0.170 ± 0.020**
Avg. Time Per Iteration (ms)	1.729 ± 0.059	**1.430 ± 0.068**	17.608 ± 0.458	4.159 ± 0.023	2.294 ± 0.123	8.016 ± 0.121	2.624 ± 0.035

algorithms. We also report the final test errors and wall-clock time per iteration in Table 3. The differences between MAML (BSGD) and MOML^{v1} in test error seem to be significant. The reason might be that MAML (BSGD) has a large optimization error due to a small batch size $K = 1, 3$. The generalization error might also contribute to the differences in test error but they are out of the scope of our paper.

Besides, it seems that MOML^{v2} performs worse than MOML^{v1} in practice. Moreover, MOML^{v2} also takes longer time per iteration than MOML^{v2} because MOML^{v2} needs to update all n personalized models while MOML^{v1} only needs to update personalized models for B sampled tasks. In our experiment, we find that the computed stochastic gradients are well bounded across the iterations such that MOML^{v1} is more appropriate.

Fitted sinusoids on five unseen tasks can be found in Appendix F.

Table 4: Test accuracy for one-shot image classification in the single-node setting.

Omniglot (20-way)							
Metrics	MAML (BSGD)	Reptile	ProtoNet	NASA	MOML^{v1}	MOML^{v2}	LocalMOML
Test Accuracy (%)	44.31 ± 1.29	46.02 ± 1.08	45.67 ± 3.30	46.15 ± 0.82	**46.35 ± 1.38**	45.81 ± 0.32	46.24 ± 1.70
CIFAR-100 (5-way)							
Metrics	MAML (BSGD)	Reptile	ProtoNet	NASA	MOML^{v1}	MOML^{v2}	LocalMOML
Test Accuracy (%)	40.18 ± 0.49	40.09 ± 0.11	31.11 ± 6.29	40.60 ± 0.33	40.49 ± 0.21	**40.82 ± 0.25**	40.38 ± 0.74

6.2 One-Shot Classification in the Single-Node Setting

We also compare our proposed MOML^{v1}, MOML^{v2} and LocalMOML with baselines NASA, MAML (BSGD), Reptile (Nichol et al., 2018), and ProtoNet (Snell et al., 2017) on the Omniglot and CIFAR-100 datasets. Among the algorithms, ProtoNet is a metric-learning-based meta-learning algorithm while the others are based on the optimization of objective Eq. (1). For the Omniglot dataset, we randomly select 25 tasks for training and 10 tasks for testing while each task has 20 classes (20 ways). For the CIFAR-100 dataset, we randomly select 17 tasks for training and 3 tasks for testing while each task has 5 classes (5 ways). There are no shared classes among the tasks. We only report the test accuracy since the loss value of ProtoNet is not directly comparable to the others. We choose the feedforward neural network with ReLU nonlinearities and 2 hidden layers of size 40 as the model. The results are averaged over 3 trials with different random seeds. As shown in Table 4, MOML variants improve the performance of MAML on average.

6.3 Image Classification in a (Simulated) Federated Learning Setting

Second, we compare our LocalMOML with baselines Per-FedAvg (Fallah et al., 2020b) and pFedMe (T. Dinh et al., 2020) on the image classification problem. We consider three data sets, MNIST, CIFAR-10, and CIFAR-100. In order to create heterogeneous data distribution, we follow the setup in Fallah et al. (2020b). In particular, For MNIST and CIFAR-10 (10 classes), we distribute the training data between $N = 50$ clients (tasks) as follows: (i) half of the clients, each has a images of each of the first five classes; (ii) The rest half clients, each has $a/2$ images from only one of the first five classes and $2a$ images from only one of the other five classes. For CIFAR-100, we consider the 20 super-classes and distribute the data similarly by dividing them into the first 10 classes and the other ten classes to run the same procedure. Similarly, we divide the test data among the clients with the same distribution as the one for the training data. We set $a = 68$ for constructing the distributed training sets of MNIST, CIFAR-10, and CIFAR-100, and set $a = 34$ for constructing the test sets of MNIST and CIFAR-10 and $a = 15$ for constructing the test sets of CIFAR-100.

We conduct experiments on four GPUs to mimic the cross-device federated learning setting, where all 50 tasks are distributed to the four GPUs roughly evenly. At each round, all four GPUs participate in the learning but each GPU only samples a batch of B' tasks from its owned tasks, and we consider two settings $B' = 1, B' = 5$. It means every round a total of $B = 4B'$ tasks are sampled for updating. We train a neural network with 2 hidden

Table 5: Comparison of final test accuracy (percentage) for 5-shot learning on three image classification data sets in a cross-device federated learning setting. The results as averaged over three runs with different random seeds.

		$H = 4$			$H = 10$		
	#Tasks (B') per-GPU	Per-FedAvg	LocalMOML	pFedMe	Per-FedAvg	LocalMOML	pFedMe
MNIST	1	**91.63 ± 0.80**	91.62 ± 0.82	91.52 ± 0.94	94.05 ± 0.37	**94.17 ± 0.44**	94.10 ± 0.21
	5	92.19 ± 0.72	**92.21 ± 0.73**	92.08 ± 0.67	94.81 ± 0.28	**94.83 ± 0.28**	94.36 ± 0.28
CIFAR-10	1	63.81 ± 0.93	**66.16 ± 1.03**	63.22 ± 1.26	66.25 ± 1.55	**68.33 ± 1.60**	64.80 ± 2.92
	5	64.44 ± 1.00	**66.87 ± 1.08**	62.68 ± 1.25	67.10 ± 2.72	**68.06 ± 1.77**	65.66 ± 2.27
CIFAR-100	1	50.65 ± 1.31	**54.00 ± 1.10**	49.64 ± 0.83	52.55 ± 1.38	**56.35 ± 0.92**	51.61 ± 1.32
	5	50.92 ± 1.31	**54.39 ± 0.78**	49.25 ± 1.06	53.37 ± 1.30	**56.45 ± 1.49**	50.95 ± 1.19

layers with each layer having 40 neurons and use the ReLU activation function. We use $\alpha = 0.001$ and the step size for the considered algorithms is tuned in a range similar to before. For all algorithms, we consider two settings of $H = 4$ and $H = 10$. The mini-batch size at every iteration (including the initial one at each round) is set to 5, that is, $K = 5, K_0 = 5$. We tune the β in a range $[0.1, 0.9]$, and run a total of 10000 iterations. For pFedMe, we tune its hyperparameter $\lambda = 100$ and set the number of steps to be 50 to solve the sub-problem accurately enough. We evaluate the test accuracy after 5-shot learning with 10 steps of fine-tuning on the test set. The final results are reported in Table 5. We can see that the proposed LocalMOML outperforms Per-FedAvg and pFedMe in almost all settings with substantial improvements on the most difficult CIFAR-100 data [3].

7. Conclusions and Discussion

In this paper, we have focused on stochastic optimization for model-agnostic meta-learning and presented two novel algorithms for both the single-node and federated learning settings. Our MOML and LocalMOML algorithms outperform existing meta-learning algorithms in several aspects. Specifically, our convergence analysis ensures that our algorithms converge to a stationary point by sampling a fixed number of tasks and a fixed number of samples per iteration. Moreover, our LocalMOML algorithm not only reduces the computational complexity but also minimizes the communication complexity compared to existing federated learning algorithms that tackle the same problem.

One limitation of the proposed MOML algorithm is that they need to maintain an individualized model for each task during the whole training process, which makes it not applicable to the problem Eq. (1) with a large/infinite number of tasks or embedded systems with small memory for learning large models. When implemented in the single-node learning setting, LocalMOML does not suffer from the same problem. It remains an interesting open problem to further improve the convergence rate of LocalMOML.

3. Note that pFedMe has a tunable hyper-parameter R: #steps to solve the inner sub-problem. Larger R leads to better performance but higher computation costs (longer running time). We choose R to make the running time of pFedMe comparable to that of Per-FedAvg/LocalMOML for a fair comparison.

Acknowledgments

This work is partially supported by NSF awards 2147253, 2110545, 1844403, and Amazon research award.

References

Ferran Alet, Maria Bauza, Kenji Kawaguchi, Nurullah Giray Kuru, Tomás Lozano-Pérez, and Leslie Kaelbling. Tailoring: encoding inductive biases by optimizing unsupervised objectives at prediction time. *Advances in Neural Information Processing Systems*, 34: 29206–29217, 2021.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

Harkirat Singh Behl, Atılım Günes Baydin, and Philip HS Torr. Alpha maml: Adaptive model-agnostic meta-learning. *arXiv preprint arXiv:1905.07435*, 2019.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *arXiv preprint arXiv:2008.10847*, 2020.

Tianyi Chen, Yuejiao Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *ArXiv*, abs/2102.04671, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

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, 2020a.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020b.

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

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*, 2018.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

S. Ghadimi, Andrzej Ruszczy'nski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim.*, 30:960–979, 2020.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. Improved schemes for episodic memory-based lifelong learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/0b5e29aa1acf8bdc5d8935d7036fa4f5-Abstract.html`.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *ArXiv*, abs/2002.05516, 2020.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *arXiv preprint arXiv:2006.09486*, 2020a.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *ArXiv*, abs/2010.07962, 2020b.

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

Yihan Jiang, Jakub Konecný, Keith Rush, and S. Kannan. Improving federated learning personalization via model agnostic meta learning. *ArXiv*, abs/1909.12488, 2019.

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. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2019.

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

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4519–4529, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks, 2016. URL `http://arxiv.org/abs/1612.00796`. cite arxiv:1612.00796.

Jakub Konecnỳ, 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.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *8th International Conference on Learning Representations (ICLR)*, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6470–6479, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Gary Marcus. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631, 2018. URL `https://arxiv.org/pdf/1801.00631.pdf`.

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 (AISTATS)*, pages 1273–1282, 2017.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *arXiv preprint arXiv:1909.04630*, 2019.

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

J. Snell, Kevin Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

Xingyou Song, W. Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. Es-maml: Simple hessian-free meta learning. *ArXiv*, abs/1910.01215, 2020.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21394–21405. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/f4f1f13c8289ac1b1ee0ff176b56fc60-Paper.pdf`.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Rasul Tutunov, Minne Li, Jun Wang, and Haitham Bou-Ammar. Compositional ADAM: an adaptive compositional solver. *CoRR*, abs/2002.03755, 2020. URL `https://arxiv.org/abs/2002.03755`.

Oriol Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, 2016.

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.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. On the global optimality of model-agnostic meta-learning. In *International Conference on Machine Learning*, pages 9837–9846. PMLR, 2020.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020.

Tianbao Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pages 629–637, 2013.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353, 2018.

Pan Zhou, X. Yuan, Huan Xu, S. Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. In *NeurIPS*, 2019.

Table 6: Notations we use throughout the paper.

Basic		
n	total number of tasks/clients	
$\mathcal{T}_i$	the i-th task	
$\mathbf{w}$	the trainable parameter	
$\ell_i(\cdot, \mathbf{z})$	loss function with the data sample $\mathbf{z}$ of i-th task	
$\mathcal{L}_i(\cdot)$	risk function for the i-th task	
$\mathbf{D}_i$	distribution of data for the i-th task	
$F_i(\cdot)$	the meta-objective of task i	Lemma 8
$F(\cdot)$	the total meta-objective	Eq. (1)
ϵ	target accuracy for an approximate stationary point	Table 1
L_i, ρ_i	Lipschitz constants of gradient $\nabla\mathcal{L}_i(\cdot)$ and Hessian $\nabla^2\mathcal{L}_i(\cdot)$	Assumption 1
σ_G^2, σ_H^2	variance upper bounds of stochastic gradient $\nabla\ell(\cdot, \mathbf{z})$ and stochastic Hessian $\nabla^2\ell(\cdot, \mathbf{z})$	Assumption 2
γ_G^2, γ_H^2	heterogeneity constants of gradients $\nabla\mathcal{L}_i(\cdot)$ and Hessian $\nabla^2\mathcal{L}_i(\cdot)$ among the tasks/clients	Assumption 5 Assumption 6
G	upper bound of the norm of gradient $\nabla\mathcal{L}_i(\cdot)$	Assumption 4
single-node Learning		
B, K	batch size of tasks and batch size of data samples per batch	
$\mathcal{B}$	size-B batch of tasks	
$\mathcal{T}_i$	the i-th task	
$\mathcal{S}_1^i, \mathcal{S}_2^i, \mathcal{S}_3^i$	independent batches of data points of task i	
$\widehat{\nabla}_{\mathcal{S}}\mathcal{L}_i(\cdot)$	stochastic gradient of $\mathcal{L}_i(\cdot)$ estimated on batch $\mathcal{S}$	
$\widehat{\Delta}_{\mathcal{B}}$	stochastic estimator of meta-gradient estimated on a batch of tasks $\mathcal{B}$	(2)
$\mathbf{u}^i$	personalized model for task i	
$\mathbf{v}_i(\mathbf{w})$	updated model by one step of gradient descent $\mathbf{v}_i(\mathbf{w}) \coloneqq \mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})$	
$\widehat{\mathbf{v}}^i$	updated model by one step of stochastic gradient descent $\widehat{\mathbf{v}}^i \coloneqq \mathbf{w} - \alpha\widehat{\nabla}\mathcal{L}_i(\mathbf{w})$	
η_t	the step size in iteration t	
β_t	the momentum factor in iteration t	
p_i	the probability of sampling task i in $\mathcal{B}_t'$	
C_p	the constant $\max_i 1/p_i - 1$	
Federated Learning		
$\mathcal{C}_i$	the i-th client	
R, H	total number of communication rounds and total number of iterations in each round	
$\Delta\mathbf{w}_r$	the average of stochastic meta-gradients over the sampled clients $i \in \mathcal{B}_r$ and local steps of round r	
$\tilde{\eta}$	the effective stepsize per round, i.e., $\tilde{\eta} \coloneqq \eta H$	

Table 7: Comparison of iteration complexities and required step sizes of algorithms under the single-node learning setting (supplementary to Table 1).

single-node Learning					
Algorithm	Stepsize	Iteration Complexity	#Task (B) Per Iteration	#Data points (K) Per Iteration	Sample Complexity
MAML (Fallah et al., 2020a)	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$	n	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}\left(n\epsilon^{-4}\right)$
SCGD (Wang et al., 2017)	$\mathcal{O}(\epsilon^{6})$	$\mathcal{O}(\epsilon^{-8})$	n	$\mathcal{O}(1)$	$\mathcal{O}(n\epsilon^{-8})$
NASA (Ghadimi et al., 2020)	$\mathcal{O}(\epsilon^{2})$	$\mathcal{O}(\epsilon^{-4})$	n	$\mathcal{O}(1)$	$\mathcal{O}(n\epsilon^{-4})$
BSGD (Hu et al., 2020)	$\mathcal{O}(\epsilon^{2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-6})$
BSpiderBoost (Hu et al., 2020)	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$	$\sqrt{n}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}\left(n\epsilon^{-2}+\sqrt{n}\epsilon^{-4}\right)$
MOML$^{\mathbf{v1}}$ (This work)	$\mathcal{O}(\epsilon^{3})$	$\mathcal{O}(n\epsilon^{-5})$	$\boldsymbol{\mathcal{O}(1)}$	$\boldsymbol{\mathcal{O}(1)}$	$\mathcal{O}\left(n\epsilon^{-5}\right)$
MOML$^{\mathbf{v2}}$ (This work)	$\mathcal{O}(\epsilon^{3})$	$\mathcal{O}(n\epsilon^{-5})$	$\boldsymbol{\mathcal{O}(1)}$	$\boldsymbol{\mathcal{O}(1)}$	$\mathcal{O}\left(n\epsilon^{-5}\right)$

Appendix A. Preliminary Lemmas

Lemma 8 (Lemma A.3 in Fallah et al. (2020a)) *If $\alpha \in [0, \frac{\sqrt{2}-1}{L})$ and Assumptions 1, 2, and 5 are satisfied. Then, for any $\mathbf{w} \in \mathbb{R}^d$, we have*

$$\|\nabla \mathcal{L}(\mathbf{w})\|^2 \leq C_1^2 \|\nabla F(\mathbf{w})\| + C_2^2 \gamma_G^2, \tag{8}$$

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla F_i(\mathbf{w})\|^2 \leq 2(1+\alpha L)^2 C_1^2 \|\nabla F(\mathbf{w})\|^2 + (1+\alpha L)^2(2C_2^2+1)\gamma_G^2, \tag{9}$$

where $L \coloneqq \max_i L_i$, $\nabla \mathcal{L}(\mathbf{w}) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathcal{L}_i(\mathbf{w})$, $F_i(\mathbf{w}) \coloneqq \mathcal{L}_i(\mathbf{w} - \alpha \nabla \mathcal{L}_i(\mathbf{w}))$, $F(\mathbf{w}) \coloneqq \frac{1}{n}\sum_{i=1}^n F_i(\mathbf{w})$, $C_1 \coloneqq \frac{1}{1-2\alpha L-\alpha^2 L^2}$, and $C_2 \coloneqq \frac{2\alpha L+\alpha^2 L^2}{1-2\alpha L-\alpha^2 L^2}$.

Lemma 9 *Assume that $X = \frac{1}{n}\sum_{i=1}^n X_i$. If we sample a size-$B$ minibatch $\mathcal{B}$ from $\{1,\ldots,n\}$ uniformly at random, we have $\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}} X_i\right] = X$ and*

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}} X_i\right\|^2\right] \leq \frac{n-B}{B(n-1)}\frac{1}{n}\sum_{i=1}^n \|X_i\|^2 + \frac{n(B-1)}{B(n-1)}\|X\|^2. \tag{10}$$

Proof First, we define random variables ξ_i and $\xi_{ii'}$ as

$$\xi_i = \begin{cases} 1 & i \in \mathcal{B} \\ 0 & i \notin \mathcal{B} \end{cases}, \quad \xi_{ii'} = \begin{cases} 1 & i \in \mathcal{B} \text{ and } i' \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

Note that $\xi_{ii'} = \xi_i \xi_{i'}$, $\mathbb{E}[\xi_i] = \Pr(i \in \mathcal{B}) = \frac{B}{n}$ and $\mathbb{E}[\xi_{ii'}] = \Pr(i \in \mathcal{B}, i' \in \mathcal{B}) = \frac{B(B-1)}{n(n-1)}$. Thus, it is clear that

$$\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}} X_i\right] = \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^n \xi_i X_i\right] = \frac{1}{B}\sum_{i=1}^n \mathbb{E}[\xi_i] X_i = \frac{1}{B}\sum_{i=1}^n \Pr(i\in\mathcal{B}) X_i = \frac{1}{B}\sum_{i=1}^n \frac{B}{n} X_i = X.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}}X_i\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{n}\xi_i X_i\right\|^2\right] = \frac{1}{B^2}\sum_{i=1}^{n}\mathbb{E}\left[\|\xi_i X_i\|^2\right] + \frac{1}{B^2}\sum_{i\neq i'}\mathbb{E}[\xi_i\xi_{i'}]\langle X_i, X_{i'}\rangle \\ &= \frac{1}{B^2}\sum_{i=1}^{n}\frac{B}{n}\|X_i\|^2 + \frac{1}{B^2}\sum_{i\neq i'}\frac{B(B-1)}{n(n-1)}\langle X_i, X_i'\rangle \\ &= \frac{n-B}{B(n-1)}\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 + \frac{n(B-1)}{B(n-1)}\|X\|^2.\end{aligned}$$

■

Lemma 10 (Corollary A.1 of Fallah et al. 2020a) *If $\alpha \in [0, 1/L]$, for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$,*

$$F(\mathbf{w}') - F(\mathbf{w}) - \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle \leq \frac{L(\mathbf{w})}{2}\left\|\mathbf{w}' - \mathbf{w}\right\|^2, \tag{11}$$

where $L(\mathbf{w}) \coloneqq 4L + \frac{2\rho\alpha}{n}\sum_{i=1}^{n}\|\nabla\mathcal{L}_i(\mathbf{w})\|$. If $\mathcal{L}_i(\cdot)$ is G-Lipschitz-continuous, that is, $\|\nabla\mathcal{L}_i(\mathbf{w})\| \leq G$ for any $\mathbf{w} \in \mathbb{R}^d$, then $F(\cdot)$ is L_F-smooth with $L_F \coloneqq 4L + 2\rho\alpha G$.

Lemma 11 (Lemma 4.4 of Fallah et al. 2020b) *If $\alpha \in (0, 1/L]$ and Assumptions 5, 4, and 6 are satisfied, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq \gamma_F^2 \coloneqq 3G^2\alpha^2\gamma_H^2 + 192\gamma_G^2. \tag{12}$$

Lemma 12 *For $\widehat{L}(\mathbf{w}) \coloneqq 4L + \frac{2\rho\alpha}{B}\sum_{i\in\mathcal{B}}\left\|\widehat{\nabla}_{\mathcal{S}}\mathcal{L}_i(\mathbf{w})\right\|$ and $\frac{(n-1)B}{n-B} \geq \lceil(8\rho\alpha\gamma_G/L)^2\rceil$, $S \geq \lceil(8\rho\alpha\sigma_G/L)^2\rceil$, we have:*

$$\frac{4}{5L(\mathbf{w})} \leq \mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right] \leq \frac{11\eta_0}{8L(\mathbf{w})} \leq \frac{\eta_0}{4L}, \qquad \frac{L(\mathbf{w})}{2}\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})^2}\right] \leq \frac{2}{L(\mathbf{w})} \leq \frac{1}{2L}, \tag{13}$$

where $L(\mathbf{w}) \coloneqq 4L + \frac{2\rho\alpha}{n}\sum_{i=1}^{n}\|\nabla\mathcal{L}_i(\mathbf{w})\|$.

Proof The results $\frac{4}{5L(\mathbf{w})} \leq \mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right]$ and $\frac{L(\mathbf{w})}{2}\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})^2}\right] \leq \frac{2}{L(\mathbf{w})} \leq \frac{1}{2L}$ can be found in Lemma 5.9 of Fallah et al. (2020a). Now we prove the rest part. Utilize Theorem A.2 of Fallah et al. (2020a) with $X = \frac{2\rho\alpha}{B'}\sum_{i\in\mathcal{B}'}\left\|\widehat{\nabla}_{\mathcal{S}'}\mathcal{L}_i(\mathbf{w})\right\|$, $c = 4L$, $k = 1$:

$$L(\mathbf{w})\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right] \leq \frac{\sigma_X^2/4L + \mu_X^2\frac{\mu_X}{\sigma_X^2+\mu_X(\mu_X+c)}}{\sigma_X^2+\mu_X^2}L(\mathbf{w}),$$

where μ_X and σ_X^2 are the mean and variance of X. Consider that $\sigma_X^2 + \mu_X(\mu_X + 4L) \geq \mu_X(\mu_X + 4L)$ and $L(\mathbf{w}) \leq \mu_X + 5L$, which is shown in (60) of Fallah et al. (2020a).

$$L(\mathbf{w})\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right] \leq \frac{\sigma_X^2/4L + \mu_X^2/(\mu_X+4L)}{\sigma_X^2+\mu_X^2}(\mu_X+5L) = \frac{\frac{5}{4}\sigma_X^2 + \frac{\mu_X^2(\mu_X+5L)}{\mu_X+4L} + \frac{\sigma_X^2\mu_X}{4L}}{\sigma_X^2+\mu_X^2}.$$

Note that $\frac{\mu_X+5L}{\mu_X+4L} = 1 + \frac{L}{\mu_X+4L} \leq \frac{5}{4}$ and $\sigma_X^2 \leq \frac{4\rho^2\alpha^2(n-B)}{B(n-1)}\left(\gamma_G^2 + \frac{\sigma_G^2}{S}\right)$

$$L(\mathbf{w})\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right] \leq \frac{5}{4} + \frac{\sigma_X^2\mu_X}{4L(\sigma_X^2+\mu_X^2)} \leq \frac{5}{4} + \frac{\sigma_X^4/L^2 + \mu_X^2}{8(\sigma_X^2+\mu_X^2)},$$

where the last inequality above uses Young's inequality. We only require $\sigma_X^2 \leq L^2$ to make $L(\mathbf{w})\mathbb{E}\left[\frac{1}{\widehat{L}(\mathbf{w})}\right] \leq \frac{11\eta_0}{8}$, which is satisfied if $\frac{(n-1)B}{n-B} \geq \lceil(8\rho\alpha\gamma_G/L)^2\rceil$, $S \geq \lceil(8\rho\alpha\sigma_G/L)^2\rceil$. ■

Appendix B. Convergence Analysis of MOML$^{\mathbf{v1}}$

Lemma 13 *For the stochastic meta-gradient estimator $\widehat{\Delta}_{\mathcal{B}_t}$ defined in (3), it holds that*

$$\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2\right] \leq C_\Delta,$$

where $C_\Delta \coloneqq (\alpha^2\sigma_H^2/K + (1+\alpha L)^2)(\sigma_G^2/K + G^2)$.

Proof Based on Assumption 1 and 4, we have

$$\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t)\right\|^2\right] = \alpha^2\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t) - \nabla^2\mathcal{L}_i(\mathbf{w}_t)\right\|^2\right] + (1+\alpha L)^2 \leq \frac{\alpha^2\sigma_H^2}{K} + (1+\alpha L)^2,$$

$$\mathbb{E}_{\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2\right] = \mathbb{E}_{\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i) - \nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2\right] + \left\|\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2 \leq \frac{\sigma_G^2}{K} + G^2.$$

Consider that $\mathcal{S}_2^i$, $\mathcal{S}_3^i$ are mutually independent.

$$\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2\right]$$
$$\leq \frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t)\right\|^2\right]\mathbb{E}_{\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2\right] \leq \left(\frac{\alpha^2\sigma_H^2}{S_2} + (1+\alpha L)^2\right)\left(\frac{\sigma_G^2}{S_3} + G^2\right).$$

■

Proof [Proof of Lemma 1] Based on the L_F-smoothness of F, we have

$$\begin{aligned}
F(\mathbf{w}_{t+1}) &\leq F(\mathbf{w}_t) + \langle\nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L_F}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2 \\
&= F(\mathbf{w}_t) - \eta\langle F(\mathbf{w}_t), \frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\rangle + \frac{\eta^2 L_F}{2}\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \\
&= F(\mathbf{w}_t) - \eta\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta\langle\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t}\rangle + \frac{\eta^2 L_F}{2}\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2.
\end{aligned}$$

Take expectation on both sides conditioned on $\mathcal{F}_t$, where $\mathcal{F}_t$ denotes all the randomness before the t-th iteration.

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) \mid \mathcal{F}_t\right] \leq F(\mathbf{w}_t) - \eta \left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta\langle \nabla F(\mathbf{w}_t), \mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\rangle + \frac{\eta^2 L_F}{2}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right]. \quad (14)$$

Consider that $\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) \mid \mathcal{F}_t\right] = \nabla F(\mathbf{w}_t)$.

$$\begin{aligned}
&\mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) - \frac{1}{B}\sum_{i\in\mathcal{B}_t}(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left((I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) - \mathbb{E}\left[(I - \alpha\widehat{\nabla}_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i) \mid \mathcal{F}_t, \mathcal{B}_t\right]\right) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t)\right)\left(\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) - \nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right) \mid \mathcal{F}_t\right].
\end{aligned}$$

Then, Young's inequality implies that

$$\begin{aligned}
&\langle \nabla F(\mathbf{w}_t), \mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\rangle \\
&= \mathbb{E}\left[\langle \nabla F(\mathbf{w}_t), \frac{1}{B}\sum_{i\in\mathcal{B}_t}\left(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t)\right)\left(\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) - \nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right)\rangle \mid \mathcal{F}_t\right] \\
&\leq \frac{\left\|\nabla F(\mathbf{w}_t)\right\|^2}{2} + \frac{(1+\alpha L)^2 L^2}{2}\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left\|\mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i\right\|^2 \mid \mathcal{F}_t\right]. \quad (15)
\end{aligned}$$

Considering (15) and setting $\alpha \leq 1/L$, the R.H.S. of (14) can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{w}_{t+1}) \mid \mathcal{F}_t\right] \\
&\leq F(\mathbf{w}_t) - \frac{\eta}{2}\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \frac{\eta^2 L_F}{2}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right] + \frac{4\eta L^2}{2}\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left\|\mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i\right\|^2 \mid \mathcal{F}_t\right].
\end{aligned}$$

Use the tower property of conditional expectation, re-arrange the terms, and unwrap the recursion above from iteration 0 to $T-1$

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \leq \frac{2F(\mathbf{w}_0)}{\eta} + \eta L_F\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2\right] + \frac{8L^2}{B}\mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{i\in\mathcal{B}_t}\left\|\mathbf{v}_i(\mathbf{w}_t) - \mathbf{u}_{t+1}^i\right\|^2\right].$$

■

Proof [Proof of Lemma 2] Based on (v1), the following equation holds

$$
\begin{aligned}
&\left\|\mathbf{v}_i(\mathbf{w}_{t_k^i}) - \mathbf{u}^i_{t^i_{k_i}+1}\right\|^2 \\
&= \left\|\mathbf{v}_i(\mathbf{w}_{t^i_{k_i}}) - (1-\beta)\mathbf{u}^i_{t^i_{k-1}+1} - \beta\widehat{\mathbf{v}}^i_{t^i_k}\right\|^2 \\
&= \mathbb{E}\left[\left\|(1-\beta)(\mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}) - \mathbf{u}^i_{t^i_{k-1}+1}) + (1-\beta)(\mathbf{v}_i(\mathbf{w}_{t^i_k}) - \mathbf{v}_i(\mathbf{w}_{t^i_{k-1}})) + \beta(\mathbf{v}_i(\mathbf{w}_{t^i_k}) - \widehat{\mathbf{v}}^i_{t^i_k})\right\|^2\right]
\end{aligned}
$$

We take expectation on both sides condition on $\mathcal{F}_{t^i_{k-1}}$

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{t_k^i}) - \mathbf{u}^i_{t^i_k+1}\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right] \\
&= \beta^2\alpha^2\mathbb{E}\left[\left\|\widehat{\nabla}_{\mathcal{S}_1^i}\mathcal{L}_i(\mathbf{w}_{t_k^i}) - \nabla\mathcal{L}_i(\mathbf{w}_{t_k^i})\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right] \\
&\qquad + (1-\beta)^2\mathbb{E}\left[\left\|(\mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}) - \mathbf{u}^i_{t^i_{k-1}+1}) + (\mathbf{v}_i(\mathbf{w}_{t^i_k}) - \mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}))\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right] \\
&\le \frac{\beta^2\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} + (1-\beta)\left\|\mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}) - \mathbf{u}^i_{t^i_{k-1}+1}\right\|^2 \\
&\qquad + (1-\beta)^2(1+1/\beta)(1+\alpha L)^2\mathbb{E}\left[\left\|\mathbf{w}_{t^i_k} - \mathbf{w}_{t^i_{k-1}}\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right], \\
&\le \frac{\beta^2\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} + (1-\beta)\left\|\mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}) - \mathbf{u}^i_{t^i_{k-1}+1}\right\|^2 + 8\eta^2\mathbb{E}\left[\left\|\sum_{\tau=t^i_{k-1}}^{t^i_k-1}\widehat{\Delta}_{\mathcal{B}_\tau}\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right]\mathbb{I}[\beta\in(0,1)],
\end{aligned}
$$

where t^i_{k-1} is the latest iteration before t^i_k that task $\mathcal{T}_i$ is also sampled, in other words, $t^i_{k-1} = \max\{\tau \mid \tau\in\mathcal{T}_i \wedge \tau < t^i_k\}$. Lemma 13 implies that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau=t^i_{k-1}}^{t^i_k-1}\widehat{\Delta}_{\mathcal{B}_\tau}\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right] &\le \mathbb{E}\left[(t^i_k - t^i_{k-1})\sum_{\tau=t^i_{k-1}}^{t^i_k-1}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_\tau}\right\|^2 \mid \mathcal{B}_\tau, \mathcal{F}_{t^i_{k-1}}\right] \mid \mathcal{F}_{t^i_{k-1}}\right] \\
&\le \mathbb{E}\left[(t^i_k - t^i_{k-1})^2 \mid \mathcal{F}_{t^i_{k-1}}\right]C_\Delta.
\end{aligned}
$$

It is worth noting that $t^i_k - t^i_{k-1}$ follows the geometric distribution. Thus, the second moment satisfies $\mathbb{E}\left[(t^i_k - t^i_{k-1})^2 \mid \mathcal{F}_{t^i_{k-1}}\right] \le \frac{2n^2}{B^2}$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{t_k^i}) - \mathbf{u}^i_{t^i_k+1}\right\|^2 \mid \mathcal{F}_{t^i_{k-1}}\right] \\
&\qquad \le (1-\beta)\left\|\mathbf{v}_i(\mathbf{w}_{t^i_{k-1}}) - \mathbf{u}^i_{t^i_{k-1}+1}\right\|^2 + \frac{16\eta^2n^2C_\Delta}{\beta B^2}\mathbb{I}[\beta\in(0,1)] + \frac{\beta^2\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}.
\end{aligned}
$$

Re-arranging the terms, summing over $k = 1, \ldots, T_i$, and using the tower property of conditional expectation leads to

$$\mathbb{E}\left[\sum_{k=0}^{T_i-1}\left\|\mathbf{v}_i(\mathbf{w}_k^i) - \mathbf{u}_{t_k^i+1}^i\right\|^2\right] \leq \left(\frac{\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{t_0^i}) - \mathbf{u}_{t_0^i+1}^i\right\|^2\right]}{\beta} + \frac{16\eta^2 n^2 C_\Delta}{\beta^2 B^2}\mathbb{E}\left[T_i\right]\right)\mathbb{I}[\beta \in (0,1)] + \frac{\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}\mathbb{E}\left[T_i\right].$$

Initializing the personalized model $\mathbf{u}^i$ as $\mathbf{u}_{t_0^i+1}^i = \mathbf{w}_{t_0^i} - \alpha\widehat{\nabla}_{\mathcal{S}_1^i}\mathcal{L}_i(\mathbf{w}_{t_0^i})$ and summing over $i = 1, \ldots, n$ results in

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=0}^{T_i-1}\left\|\mathbf{v}_i(\mathbf{w}_k^i) - \mathbf{u}_{t_k^i+1}^i\right\|^2\right] \leq \frac{n\sigma_G^2}{\beta|\mathcal{S}_1^i|}\mathbb{I}[\beta \in (0,1)] + \left(\frac{16\eta^2 n^2 C_\Delta}{\beta^2 B^2}\mathbb{I}[\beta \in (0,1)] + \frac{\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}\right)\mathbb{E}\left[\sum_{i=1}^{n} T_i\right].$$

Note that $\sum_{i=1}^n T_i = T$ based on the definition. ■

Theorem 14 (Detailed Version of Theorem 3) *Under Assumptions 1, 2 3, 4,* MOML$^{\text{v1}}$ *with stepsizes* $\eta_t = \frac{B^{2/5}}{n^{2/5}T^{3/5}}$, $\beta_t = \frac{n^{2/5}}{B^{2/5}T^{2/5}} < 1$ *and constant batch sizes* $|\mathcal{S}_1^i| = |\mathcal{S}_2^i| = |\mathcal{S}_3^i| = K = 1$, $|\mathcal{B}_t| = B = 1$ *can find a stationary point* $\mathbf{w}_\tau$ *in* $T = \mathcal{O}(n\epsilon^{-5})$ *iterations.*

Proof Based on Lemma 1 and Lemma 2, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \leq \frac{2F(\mathbf{w}_0)}{\eta T} + \eta L_F C_\Delta + \frac{8L^2}{B}\left(\frac{n\sigma_G^2}{\beta K T} + \frac{16\eta^2 n^2 C_\Delta}{\beta^2 B^2} + \frac{\beta\alpha^2\sigma_G^2}{K}\right).$$

Choosing $B = 1$, $K = 1$, $\eta = \frac{B^{2/5}}{n^{2/5}T^{3/5}}$ and $\beta = \frac{n^{2/5}}{B^{2/5}T^{2/5}} < 1$ leads to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \leq \frac{2n^{2/5}F(\mathbf{w}_0)}{T^{2/5}} + \frac{L_F C_\Delta}{n^{2/5}T^{3/5}} + \frac{8L^2\sigma_G^2 n^{3/5}}{T^{3/5}} + \frac{144L^2 C_\Delta n^{2/5}}{T^{2/5}} + \frac{8L^2\alpha^2\sigma_G^2 n^{2/5}}{T^{2/5}}.$$

For $\mathbf{w}_\tau$ and τ is sampled from $0, \ldots, T-1$ uniformly at random, $\mathbb{E}\left[\|\nabla F(\mathbf{w}_\tau)\|^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right]$. Making the R.H.S. of the upper bound of $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right]$ be equal to or smaller than ϵ^2 finishes the proof. ■

Appendix C. Convergence Analysis of MOML$^{\mathbf{v2}}$

Lemma 15 *For the stochastic estimator* $\widehat{\Delta}_{\mathcal{B}_t} = \frac{1}{B}\sum_{i\in\mathcal{B}_t}(I-\alpha\nabla^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\nabla_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)$ *of the meta-gradient, we have*

$$\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right] \le C_3\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right] + C_4\left\|\nabla F(\mathbf{w}_t)\right\|^2 + C_5, \tag{16}$$

where $\mathcal{F}_t$ *denotes all randomness occurred before the* t*-th iteration, and the constants are defined as* $C_3 \coloneqq 2L^2\left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)$, $C_4 \coloneqq 4\left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)(1+\alpha L)^2C_1$, *and* $C_5 \coloneqq \left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)\left(\frac{\sigma_G^2}{|\mathcal{S}_3^i|} + 2(1+\alpha L)^2(2C_2^2+1)\gamma_G^2\right)$.

Proof The definition of the stochastic estimator $\widehat{\Delta}_{\mathcal{B}_t}$ implies that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I-\alpha\nabla^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\nabla_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left((I-\alpha\nabla^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\nabla_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{t+1}^i) - (I-\alpha\nabla^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right)\right\|^2 \mid \mathcal{F}_t\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\left((I-\alpha\nabla^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i) - (I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right)\right\|^2 \mid \mathcal{F}_t\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2 \mid \mathcal{F}_t\right] \\
&\le \left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)\left(\frac{\sigma_G^2}{|\mathcal{S}_3^i|} + \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)\right\|^2 \mid \mathcal{F}_t\right]\right) \\
&\le \left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)\left(\frac{\sigma_G^2}{|\mathcal{S}_3^i|} + \frac{2}{n}\sum_{i=1}^n\left\|\nabla F_i(\mathbf{w}_t))\right\|^2 + \frac{2L^2}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right]\right),
\end{aligned}$$

where the last inequality uses the fact $\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) = \nabla F_i(\mathbf{w}_t)$. Lemma 8 shows that

$$\frac{1}{n}\sum_{i=1}^n\left\|\nabla F_i(\mathbf{w}_t)\right\|^2 \le 2(1+\alpha L)^2C_1\left\|\nabla F(\mathbf{w}_t)\right\|^2 + (1+\alpha L)^2(2C_2^2+1)\gamma_G^2.$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right] &= \left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)\left(\frac{\sigma_G^2}{|\mathcal{S}_3^i|} + 2(1+\alpha L)^2(2C_2^2+1)\gamma_G^2\right) \\
&\quad + 4\left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)(1+\alpha L)^2C_1\left\|\nabla F(\mathbf{w}_t)\right\|^2 \\
&\quad + 2L^2\left((1+\alpha L)^2 + \frac{\alpha^2\sigma_H^2}{|\mathcal{S}_2^i|}\right)\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right].
\end{aligned}$$

■

Apart from the notations in Table 6, we define that $\mathbf{I}_i \coloneqq (0_{d\times d}, \ldots, I_{d\times d}, \ldots, 0_{d\times d})^\top \in \mathbb{R}^{nd\times d}$ (where the i-th block in $\mathbf{I}_i$ is an identity matrix while the others are zeros), $\bar{\mathbf{w}}_t \coloneqq \left(\mathbf{w}_t^\top, \ldots, \mathbf{w}_t^\top\right)^\top \in \mathbb{R}^{nd}$, $\mathbf{u}_t = \left([\mathbf{u}_t^1]^\top, \ldots, [\mathbf{u}_t^n]^\top\right)^\top \in \mathbb{R}^{nd}$, $\widehat{\mathbf{g}}_t \coloneqq \sum_{i\in\mathcal{B}_t'} \frac{1}{\alpha p_i}\mathbf{I}_i\left(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i\right))$, $\bar{\mathbf{g}}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i))/\alpha$, $\mathbf{g}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \mathbf{v}_i(\mathbf{w}_t))/\alpha$, $\widetilde{\mathbf{w}}_t \coloneqq \bar{\mathbf{w}}_t - \alpha\mathbf{g}_t$. Then, we can re-write the update rule of the personalized models $\mathbf{u}_t^i$ for all tasks $i \in [n]$ in a more succinct expression $\mathbf{u}_{t+1} = (1-\beta_t)\mathbf{u}_t + \beta_t\left(\bar{\mathbf{w}}_t - \alpha\widehat{\mathbf{g}}_t\right)$.

Lemma 16 *For* $\widehat{\mathbf{g}}_t \coloneqq \sum_{i\in\mathcal{B}_t'} \frac{1}{\alpha p_i}\mathbf{I}_i\left(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i\right))$ *and* $\mathbf{g}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \mathbf{v}_i(\mathbf{w}_t))/\alpha$, *we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] &\le 2nC_pC_1^2\left\|\nabla F(\mathbf{w}_{t-1})\right\|^2 \\ &\quad + nC_p\left((2C_2^2+1)\gamma_G^2 + \frac{2\sigma_G^2}{|\mathcal{S}_1^i|}\right) + 2nLC_p\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2, \end{aligned} \tag{17}$$

where $C_p \coloneqq \max_i\left(\frac{1}{p_i} - 1\right)$.

Proof Consider that $\widehat{\mathbf{g}}_t \coloneqq \sum_{i\in\mathcal{B}_t'} \frac{1}{\alpha p_i}\mathbf{I}_i\left(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i\right))$, $\bar{\mathbf{g}}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i))/\alpha$, $\mathbf{g}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \mathbf{v}_i(\mathbf{w}_t)))/\alpha$.

$$\mathbb{E}\left[\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{g}}_t - \bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] + \mathbb{E}\left[\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right]. \tag{18}$$

The first term on the right hand side of (18) can be upper bounded as

$$\mathbb{E}\left[\left\|\widehat{\mathbf{g}}_t - \bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] \le C_p\mathbb{E}\left[\left\|\bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] = C_p\left(\left\|\mathbf{g}_t\right\|^2 + \mathbb{E}\left[\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right]\right), \tag{19}$$

where we define $C_p \coloneqq \max_i\left(\frac{1}{p_i} - 1\right)$. Note that $\left\|\mathbf{g}_t\right\|^2 = \sum_{i=1}^n \left\|\nabla\mathcal{L}_i(\mathbf{w}_t)\right\|^2$. Then,

$$\begin{aligned}\left\|\mathbf{g}_t\right\|^2 &= n\left\|\nabla\mathcal{L}(\mathbf{w}_t)\right\|^2 + n\gamma_G^2 \le 2n\left\|\nabla\mathcal{L}(\mathbf{w}_{t-1})\right\|^2 + 2nL\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2 + n\gamma_G^2 \\ &\le 2nC_1^2\left\|\nabla F(\mathbf{w}_{t-1})\right\|^2 + n(2C_2^2+1)\gamma_G^2 + 2nL\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2. \end{aligned} \tag{20}$$

The last inequality above utilizes Lemma 8. Besides,

$$\mathbb{E}\left[\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] = \sum_{i=1}^n \mathbb{E}\left[\left\|\widehat{\nabla}_{\mathcal{S}_1^i}\mathcal{L}_i(\mathbf{w}_t) - \nabla\mathcal{L}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right] \le \frac{n\sigma_G^2}{|\mathcal{S}_1^i|}. \tag{21}$$

According to (18), (19), (20), and (21), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|^2 \mid \mathcal{F}_t\right] &\le 2nC_pC_1^2\left\|\nabla F(\mathbf{w}_{t-1})\right\|^2 \\ &\quad + nC_p\left((2C_2^2+1)\gamma_G^2 + \frac{2\sigma_G^2}{|\mathcal{S}_1^i|}\right) + 2nLC_p\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2. \end{aligned}$$

■

Proof [Proof of Lemma 4] Recall that the notations $\mathbf{v}_i(\mathbf{w}_t) \coloneqq \mathbf{w}_t - \alpha\nabla\mathcal{L}_i(\mathbf{w}_t)$, $\bar{\mathbf{w}}_t \coloneqq \left(\mathbf{w}_t^\top, \ldots, \mathbf{w}_t^\top\right)^\top \in \mathbb{R}^{nd}$, $\mathbf{u}_t = \left([\mathbf{u}_t^1]^\top, \ldots, [\mathbf{u}_t^n]^\top\right)^\top$, $\widehat{\mathbf{g}}_t \coloneqq \sum_{i\in\mathcal{B}_t'} \frac{1}{\alpha p_i}\mathbf{I}_i\left(\mathbf{w}_t - \widehat{\mathbf{v}}_t^i\right)\big)$, $\mathbf{g}_t \coloneqq \sum_{i=1}^n \mathbf{I}_i(\mathbf{w}_t - \mathbf{v}_i(\mathbf{w}_t))/\alpha$, $\widetilde{\mathbf{w}}_t \coloneqq \bar{\mathbf{w}}_t - \alpha\mathbf{g}_t$. The update rule of the personalized models is $\mathbf{u}_{t+1} = (1-\beta_t)\mathbf{u}_t + \beta_t\left(\bar{\mathbf{w}}_t - \alpha\widehat{\mathbf{g}}_t\right)$. We define that $\Upsilon_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2$.

$$
\begin{aligned}
&\mathbb{E}\left[\Upsilon_{t+1} \mid \mathcal{F}_{t+1}\right] = \frac{1}{n}\mathbb{E}\left[\left\|\mathbf{u}_{t+2} - \widetilde{\mathbf{w}}_{t+1}\right\|^2 \mid \mathcal{F}_{t+1}\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left\|\widetilde{\mathbf{w}}_{t+1} - (1-\beta_{t+1})\mathbf{u}_{t+1} + \beta_{t+1}\left(\bar{\mathbf{w}}_{t+1} - \alpha\widehat{\mathbf{g}}_{t+1}\right)\right\|^2 \mid \mathcal{F}_{t+1}\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left\|(1-\beta_{t+1})(\widetilde{\mathbf{w}}_t - \mathbf{u}_{t+1}) + (1-\beta_{t+1})\left(\widetilde{\mathbf{w}}_{t+1} - \widetilde{\mathbf{w}}_t\right) + \alpha\beta_{t+1}(\mathbf{g}_{t+1} - \widehat{\mathbf{g}}_{t+1})\right\|^2 \mid \mathcal{F}_{t+1}\right] \\
&\le (1-\beta_{t+1})\frac{1}{n}\left\|\widetilde{\mathbf{w}}_t - \mathbf{u}_{t+1}\right\|^2 + \frac{8(1+\alpha L)^2}{\beta_{t+1}n}\left\|\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t\right\|^2 + \frac{\alpha^2\beta_{t+1}^2}{n}\mathbb{E}\left[\left\|\mathbf{g}_{t+1} - \widehat{\mathbf{g}}_{t+1}\right\|^2 \mid \mathcal{F}_{t+1}\right] \\
&= (1-\beta_{t+1})\frac{1}{n}\sum_{i=1}^n\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 + \frac{8(1+\alpha L)^2}{\beta_{t+1}}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2 + \frac{\alpha^2\beta_{t+1}^2}{n}\mathbb{E}\left[\left\|\mathbf{g}_{t+1} - \widehat{\mathbf{g}}_{t+1}\right\|^2 \mid \mathcal{F}_{t+1}\right].
\end{aligned}
$$

Based on Lemma 16, we have

$$
\begin{aligned}
\mathbb{E}\left[\Upsilon_{t+1} \mid \mathcal{F}_{t+1}\right] \le{}& (1-\beta_{t+1})\Upsilon_t + 2\left(\frac{4(1+\alpha L)^2}{\beta_{t+1}} + \beta_{t+1}^2\alpha^2 L C_p\right)\eta_t^2\left\|\Delta_{\mathcal{B}_t}\right\|^2 \\
&+ 2\beta_{t+1}^2\alpha^2 C_p C_1^2\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \beta_{t+1}^2\alpha^2 C_p(2C_2^2+1)\gamma_G^2 + \frac{2\beta_{t+1}^2\alpha^2 C_p\sigma_G^2}{|\mathcal{S}_1^i|}.
\end{aligned}
$$

We choose $\beta_{t+1} = 6L^2\eta_0^{-1/3}\eta_t$. Since we need to ensure $\beta_t \le 1$ for any t, we only need to maintain $\eta_0 \le \left(\frac{2}{3L}\right)^{\frac{3}{2}}$. Then, $\beta_{t+1}^2 = 36L^4\eta_0^{-2/3}\eta_t^2 \le 3L^2\eta_0^{\frac{4}{3}}$.

$$
\mathbb{E}\left[\Upsilon_{t+1} \mid \mathcal{F}_{t+1}\right] \le (1 - 6L^2\eta_0^{-1/3}\eta_t)\Upsilon_t + C_6\eta_0^{1/3}\eta_t\left\|\Delta_{\mathcal{B}_t}\right\|^2 + \eta_0^{1/3}\eta_t C_7\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_0^{4/3}C_8,
$$

where $C_6 \coloneqq \frac{4(1+\alpha L)^2+\alpha^2 L C_p}{3L^2}$, $C_7 \coloneqq 18L^3\alpha^2 C_p C_1^2$, $C_8 \coloneqq 3L^2\left(\alpha^2 C_p(2C_2^2+1)\gamma_G^2 + \frac{2\alpha^2 C_p\sigma_G^2}{|\mathcal{S}_1^i|}\right)$. Then, in view of the tower property of conditional expectation, we have

$$
\begin{aligned}
\mathbb{E}\left[\Upsilon_{t+1} \mid \mathcal{F}_t\right] \le{}& (1 - 6L^2\eta_0^{-1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right])\mathbb{E}\left[\Upsilon_t \mid \mathcal{F}_t\right] \\
&+ \eta_0^{1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]C_7\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_0^{4/3}C_8 \\
&+ C_6\eta_0^{1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]\left(C_3\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\Upsilon_t \mid \mathcal{F}_t\right] + C_4\left\|\nabla F(\mathbf{w}_t)\right\|^2 + C_5\right) \\
={}& (1 - 6L^2\eta_0^{-1/3}(1 - C_3C_6\eta_0^{2/3}/6L^2)\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right])\mathbb{E}\left[\Upsilon_t \mid \mathcal{F}_t\right] \\
&+ \eta_0^{1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right](C_7 + C_4C_6)\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_0^{4/3}C_8 + \eta_0^{1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]C_5C_6 \\
\le{}& \left(1 - 3L^2\eta_0^{-1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]\right)\mathbb{E}\left[\Upsilon_t \mid \mathcal{F}_t\right] \\
&+ \eta_0^{1/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right](C_7 + C_4C_6)\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_0^{4/3}\left(C_8 + \frac{C_5C_6}{4L}\right), \qquad (22)
\end{aligned}
$$

where the last step holds when $\eta_0 \le \left(\frac{3L^2}{C_3C_6}\right)^{3/2}$. We define $C_9 \coloneqq C_7 + C_4C_6$, $C_{10} \coloneqq C_8 + C_5C_6/(4L)$. ■

Lemma 17 *If we set* $\eta_0 \leq \min\left\{\frac{2L^2}{5C_3}, \frac{1}{8C_9^{3/2}}, \frac{1}{20C_4}\right\}$ *and define the potential function* Φ_t *as* $\Phi_t \coloneqq \eta_0^{1/3}\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 + F(\mathbf{w}_t)$*, we have*

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq \mathbb{E}\left[\Phi_t\right] - \frac{\eta_0}{80}\min\left\{\frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}{L+\rho\alpha\sigma}, \frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|\right]}{\rho\alpha}\right\} + \eta_0^{5/3}\left(\frac{\eta_0^{1/3}C_5}{2} + C_{10}\right).$$

Proof Based on Lemma 10, we have

$$\begin{aligned}
F(\mathbf{w}_{t+1}) &\leq F(\mathbf{w}_t) + \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L(\mathbf{w}_t)}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2 \\
&= F(\mathbf{w}_t) - \eta_t\langle F(\mathbf{w}_t), \widehat{\Delta}_{\mathcal{B}_t}\rangle + \frac{\eta_t^2 L(\mathbf{w}_t)}{2}\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \\
&= F(\mathbf{w}_t) - \eta_t\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \eta_t\langle \nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t}\rangle + \frac{\eta_t^2 L(\mathbf{w}_t)}{2}\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2.
\end{aligned}$$

Consider the step size η_t and $\widehat{\Delta}_{\mathcal{B}_t}$ are independent. Take expectation on both sides conditioned on $\mathcal{F}_t$, where $\mathcal{F}_t$ denotes all randomness occurred before the t-th iteration.

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{w}_{t+1}) \mid \mathcal{F}_t\right] \\
&\leq F(\mathbf{w}_t) - \mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]\left\|\nabla F(\mathbf{w}_t)\right\|^2 \\
&\quad + \mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]\langle \nabla F(\mathbf{w}_t), \mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\rangle + \frac{\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right] L(\mathbf{w}_t)}{2}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right]
\end{aligned}$$

Consider the fact $\mathbb{E}\left[\widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))\nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i) \mid \mathcal{F}_t\right]$.

$$\mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_t))(\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_t)) - \nabla\mathcal{L}_i(\mathbf{u}_{t+1}^i)) \mid \mathcal{F}_t\right]$$

Since $\|\cdot\|$ is a convex function, we have the following equation based on the Jensen's and Cauchy-Schwarz inequalities

$$\begin{aligned}
&\langle \nabla F(\mathbf{w}_t), \mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\rangle \leq \left\|\nabla F(\mathbf{w}_t)\right\|\left\|\mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\right\| \\
&\leq \frac{\left\|\nabla F(\mathbf{w}_t)\right\|^2}{2} + \frac{\left\|\mathbb{E}\left[\nabla F(\mathbf{w}_t) - \widehat{\Delta}_{\mathcal{B}_t} \mid \mathcal{F}_t\right]\right\|^2}{2} \\
&\leq \frac{\left\|\nabla F(\mathbf{w}_t)\right\|^2}{2} + \frac{(1+\alpha L)^2 L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right]}{2} \\
&\leq \frac{\left\|\nabla F(\mathbf{w}_t)\right\|^2}{2} + \frac{4L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right]}{2},
\end{aligned}$$

where the last inequality holds when $\alpha \leq 1/L$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{w}_{t+1}) \mid \mathcal{F}_t\right] \\
&\leq F(\mathbf{w}_t) - \frac{\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]}{2}\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \frac{4L^2\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right]}{2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right] \\
&\quad + \frac{L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]}{2}\mathbb{E}\left[\left\|\widehat{\Delta}_{\mathcal{B}_t}\right\|^2 \mid \mathcal{F}_t\right] \\
&\leq F(\mathbf{w}_t) - \frac{\left(\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] - C_4 L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]\right)}{2}\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \frac{C_5 L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]}{2} \\
&\quad + \frac{\left(4L^2\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] + L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]C_3\right)}{2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 \mid \mathcal{F}_t\right]. \qquad (23)
\end{aligned}$$

Based on Lemma 12, we can derive that

$$-3L^2\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] + \frac{\left(4L^2\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] + L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]C_3\right)}{2} = -\frac{4L^2\eta_0}{5L(\mathbf{w}_t)} + \frac{2C_3\eta_0^2}{L(\mathbf{w}_t)} \leq 0,$$

where we need $\eta_0 \leq \frac{2L^2}{5C_3}$. Besides, if $\eta_0 \leq \frac{1}{8C_9^{3/2}}$ and $\eta_0 \leq \frac{1}{20C_4}$, we have

$$C_9\eta_0^{2/3}\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] - \frac{\left(\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] - C_4 L(\mathbf{w}_t)\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]\right)}{2} \leq -\frac{\eta_0}{5L(\mathbf{w}_t)} + \frac{2\eta_0^2 C_4}{L(\mathbf{w}_t)} \leq -\frac{\eta_0}{10L(\mathbf{w}_t)}.$$

Multiplying (22) by $\eta_0^{1/3}$ and summing it to (23) leads to

$$\begin{aligned}
&\eta_0^{1/3}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{u}_{t+2}^i - \mathbf{v}_i(\mathbf{w}_{t+1})\right\|^2\right] + \mathbb{E}\left[F(\mathbf{w}_{t+1})\right] \\
&\leq \eta_0^{1/3}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2\right] + \mathbb{E}\left[F(\mathbf{w}_t)\right] - \mathbb{E}\left[\frac{\eta_0}{10L(\mathbf{w}_t)}\left\|\nabla F(\mathbf{w}_t)\right\|^2\right] + \eta_0^{5/3}\left(\eta_0^{1/3}C_5/2 + C_{10}\right).
\end{aligned}$$

Define that $\Phi_t \coloneqq \eta_0^{1/3}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{u}_{t+1}^i - \mathbf{v}_i(\mathbf{w}_t)\right\|^2 + F(\mathbf{w}_t)$. Besides, utilize (103) $\sim$ (106) in Fallah et al. (2020a):

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq \mathbb{E}\left[\Phi_t\right] - \frac{\eta_0}{80}\min\left\{\frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}{L + \rho\alpha\sigma}, \frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|\right]}{\rho\alpha}\right\} + \eta_0^{5/3}\left(\eta_0^{1/3}C_5/2 + C_{10}\right).$$

■

Theorem 18 (Detailed Version of Thoerem 5) *Under Assumptions 1, 2, and 5, it is guaranteed that* MOML$^{\text{v2}}$ *can find an ϵ-stationary point in $\frac{160(L+\rho\alpha(\sigma+\epsilon))\Phi_0}{C_{11}\epsilon^5}$ iterations, where $C_{11} = \Omega(1/C_p)$, $C_p = \max_i 1/p_i - 1$.*

Proof Based on Lemma 17, we can derive that:

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq \mathbb{E}\left[\Phi_t\right] - \frac{\eta_0}{80}\min\left\{\frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}{L+\rho\alpha\sigma}, \frac{\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|\right]}{\rho\alpha}\right\} + \eta_0^{5/3}\left(\eta_0^{1/3}C_5/2 + C_{10}\right).$$

Suppose that $\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|\right] \geq \epsilon$ and $\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \geq \left(\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_t)\right\|\right]\right)^2 \geq \epsilon^2, \forall t \in 1, \ldots, T$. Otherwise, we can find an ϵ-stationary point in the first T iterations. Thus, choosing $\eta_0 = C_{11}\epsilon^3$ ($C_{11} > 0$) and telescoping over the T iterations leads to:

$$T\frac{\epsilon^2}{L+\rho\alpha(\sigma+\epsilon)} \leq T\min\left\{\frac{\epsilon^2}{L+\rho\alpha\sigma}, \frac{\epsilon}{\rho\alpha}\right\} \leq \frac{80\Phi_0}{C_{11}\epsilon^3} + 80TC_{11}^{2/3}\epsilon^2\left(C_9^{1/3}\epsilon C_5/2 + C_{10}\right),$$

where $C_{11} \coloneqq \min\left\{\frac{1}{160C_5(L+\rho\alpha(\sigma+\epsilon))}, \frac{1}{(320C_9(L+\rho\alpha(\sigma+\epsilon)))^{3/2}}\right\}$. Note that $\eta_0 = C_{11}\epsilon^3$ can satisfy the requirements on η_0 in Lemma 4 and Lemma 17. Thus, we can find at least one ϵ-stationary point if $T \geq \frac{160(L+\rho\alpha(\sigma+\epsilon))\Phi_0}{C_{11}\epsilon^5}$. ■

Appendix D. Convergence Analysis of LocalMOML

To tackle with partial client sampling, we follow the ideas of Li et al. (2019); Karimireddy et al. (2020): in each round, the global model $\mathbf{w}_r$ is sent to the sampled clients $i \in \mathcal{B}_r$ and the sampled clients run local step. Here we assume that $\mathbf{w}_r$ is also *virtually* sent to the other clients $i \notin \mathcal{B}_r$ and those clients also *virtually* run local step. After H iterations, only $\Delta\mathbf{w}_r^i$, $i \in \mathcal{B}_r$ are aggregated to compute $\mathbf{w}_{r+1}$. It is worth noting that the extra communication of $\mathbf{w}_r$ to clients $i \notin \mathcal{B}_r$ and the local steps on clients $i \notin \mathcal{B}_r$ are only used in the proof and not actually executed when running Algorithm 3. The lemma below is key to our analysis.

Lemma 19 *After one round of* LocalMOML*, it satisfies that:*

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{r+1})\right] \leq{}& \mathbb{E}\left[F(\mathbf{w}_r)\right] - \frac{\tilde{\eta}}{2}\left(1 - 8\tilde{\eta} - \frac{8\tilde{\eta}(n-B)}{B(n-1)}\right)\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right] \\ &+ \frac{\tilde{\eta}}{2}\left(C_{\rho,G,L} + 8\tilde{\eta}L_F^2 + \frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\right\|^2\right] \\ &+ 8\tilde{\eta}\left(1 + L^2\tilde{\eta}\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2\right] + \frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2 + \frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}. \end{aligned} \tag{24}$$

where $\hat{\sigma}^2 \coloneqq \frac{2\sigma_G^2}{S_3} + \frac{2\alpha^2\sigma_G^2}{S_3}\left(\frac{\sigma_H^2}{S_2} + L^2\right) + \frac{\alpha^2G^2\sigma_H^2}{S_2}$, $\tilde{\eta} = \eta H$, $C_{\rho,G,L} \coloneqq 2G^2\rho^2/L^2 + 16L^2$.

Proof Based on the smoothness of F shown in Lemma 10, we have

$$F(\mathbf{w}_{r+1}) \leq F(\mathbf{w}_r) - \langle\nabla F(\mathbf{w}_r), \Delta\mathbf{w}_r\rangle + \frac{L_F}{2}\left\|\Delta\mathbf{w}_r\right\|^2.$$

Let $\tilde{\eta} \coloneqq H\eta$ and $\mathcal{F}_r$ denotes all randomness occurred before the communication round r. Note that $\Delta \mathbf{w}_r = \frac{\tilde{\eta}}{BH} \sum_{i \in \mathcal{B}_r} \sum_{h=1}^{H} \widehat{\Delta}_{r,h}^i$ and $\mathbb{E}\left[\Delta \mathbf{w}_r\right] = \frac{\tilde{\eta}}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{E}\left[\widehat{\Delta}_{r,h}^i\right]$ because $\widehat{\Delta}_{r,h}^i$ does not depend on the client sampling $\mathcal{B}_r$.

$$\mathbb{E}\left[F(\mathbf{w}_{r+1}) \mid \mathcal{F}_r\right] \leq F(\mathbf{w}_r) - \tilde{\eta} \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \langle \nabla F(\mathbf{w}_r), \mathbb{E}\left[\widehat{\Delta}_{r,h}^i \mid \mathcal{F}_r\right] \rangle + \mathbb{E}\left[\left\|\Delta \mathbf{w}_r\right\|^2 \mid \mathcal{F}_r\right]. \tag{25}$$

The second term on the right hand side can be decomposed as

$$\begin{aligned}
&- \tilde{\eta} \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \langle \nabla F(\mathbf{w}_r), \mathbb{E}\left[\widehat{\Delta}_{r,h}^i \mid \mathcal{F}_r\right] \rangle \\
&= -\tilde{\eta} \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \langle \nabla F(\mathbf{w}_r), \mathbb{E}\left[\widehat{\Delta}_{r,h}^i - \nabla F(\mathbf{w}_r) \mid \mathcal{F}_r\right] \rangle - \tilde{\eta} \left\|\nabla F(\mathbf{w}_r)\right\|^2.
\end{aligned}$$

Based on the definition of $\widehat{\Delta}_{r,h}^i$, we have

$$\begin{aligned}
&- \tilde{\eta} \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \langle \nabla F(\mathbf{w}_r), \mathbb{E}\left[\widehat{\Delta}_{r,h}^i - \nabla F(\mathbf{w}_r) \mid \mathcal{F}_r\right] \rangle \\
&= -\tilde{\eta} \mathbb{E}\left[\langle \nabla F(\mathbf{w}_r), \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \left((I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_r))\right) \rangle \mid \mathcal{F}_r\right] \\
&\leq \frac{\tilde{\eta}}{2} \left\|\nabla F(\mathbf{w}_r)\right\|^2 \\
&\quad + \frac{\tilde{\eta}}{2} \mathbb{E}\left[\left\| \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \left((I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_r))\right) \right\|^2 \mid \mathcal{F}_r\right].
\end{aligned}$$

The second term on the right hand side can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \left((I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_r))\right) \right\|^2 \mid \mathcal{F}_r\right] \\
&\leq 2\mathbb{E}\left[\left\| \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \left((I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i)\right) \right\|^2 \mid \mathcal{F}_r\right] \\
&\quad + 2\mathbb{E}\left[\left\| \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \left((I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha \nabla^2 \mathcal{L}_i(\mathbf{w}_r)) \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_r))\right) \right\|^2 \mid \mathcal{F}_r\right] \\
&\leq 2\alpha^2 \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{E}\left[\left\|\nabla^2 \mathcal{L}_i(\mathbf{w}_{r,h}^i) - \nabla^2 \mathcal{L}_i(\mathbf{w}_r)\right\|^2 \left\|\nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2 \mid \mathcal{F}_r\right] \\
&\quad + 4 \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{E}\left[\left\|\nabla \mathcal{L}_i(\mathbf{u}_{r,h}^i) - \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2 \left\|I - \alpha \nabla \mathcal{L}_i(\mathbf{w}_r)\right\|^2 \mid \mathcal{F}_r\right] \\
&\quad + 4 \frac{1}{nH} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{E}\left[\left\|\nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i)) - \nabla \mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_r))\right\|^2 \left\|I - \alpha \nabla \mathcal{L}_i(\mathbf{w}_r)\right\|^2 \mid \mathcal{F}_r\right].
\end{aligned}$$

Based on Assumption 1, 4, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\left((I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_r))\nabla\mathcal{L}_i(\mathbf{w}_r)\right)\right\|^2 \mid \mathcal{F}_r\right]\\
&\le 2\alpha^2G^2\rho^2\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2 \mid \mathcal{F}_r\right]+4(1+\alpha L)^2\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^i-\mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2 \mid \mathcal{F}_r\right]\\
&\quad+4(1+\alpha L)^2L^2\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2 \mid \mathcal{F}_r\right]\\
&= C_{\rho,G,L}\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2 \mid \mathcal{F}_r\right]+16\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^i-\mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2 \mid \mathcal{F}_r\right],
\end{aligned}$$

where $C_{\rho,G,L}\coloneqq 2\alpha^2G^2\rho^2+4(1+\alpha L)^2L^2\le 2G^2\rho^2/L^2+16L^2$ when $\alpha\le 1/L$. Besides, the last term on the right hand side of (25) can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}\widehat{\Delta}_{r,h}^i\right\|^2 \mid \mathcal{F}_r\right]\\
&\le \mathbb{E}\left[\frac{1}{B^2H^2}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\widehat{\Delta}_{r,h}^i-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right] \mid \mathcal{F}_r\right]\\
&\quad+\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right],
\end{aligned}\tag{26}$$

which is due to $\widehat{\Delta}_{r,h}^i=(I-\alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)$ and

$$\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[(I-\alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right]=0.$$

Next, we can decompose $\widehat{\Delta}_{r,h}^i-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_r))\nabla\mathcal{L}_i(\mathbf{w}_r)$ as

$$\begin{aligned}
&(I-\alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\\
&=\left(\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right)-\alpha\left(\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right)\\
&\quad-\alpha\left(\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right),
\end{aligned}$$

Then, the term $\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\widehat{\Delta}_{r,h}^i-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]$ can be upper bounded as

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\widehat{\Delta}_{r,h}^i-(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]\\
&=\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)+\alpha(\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i))\right\|^2\right]\\
&\quad+\alpha^2\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|\left(\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)-\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right)\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]\\
&\le 2\mathbb{E}_{\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]+\alpha^2G^2\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)-\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right\|^2\right]\\
&\quad+2\alpha^2\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right\|^2\right]\mathbb{E}_{\mathcal{S}_3^i}\left[\left\|\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)-\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]\\
&\le\frac{2\sigma_G^2}{|\mathcal{S}_3^i|}+\frac{2\alpha^2\sigma_G^2}{|\mathcal{S}_3^i|}\mathbb{E}_{\mathcal{S}_2^i}\left[\left\|\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i)-\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right\|^2+L^2\right]+\frac{\alpha^2G^2\sigma_H^2}{|\mathcal{S}_2^i|}\\
&\le\frac{2\sigma_G^2}{|\mathcal{S}_3^i|}+\frac{2\alpha^2\sigma_G^2}{|\mathcal{S}_3^i|}\left(\frac{\sigma_H^2}{|\mathcal{S}_2^i|}+L^2\right)+\frac{\alpha^2G^2\sigma_H^2}{|\mathcal{S}_2^i|}\coloneqq\hat{\sigma}^2.
\end{aligned}
$$

Besides, the last term on the right hand side of (26) can be upper bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]\\
&\le 2\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2\mid\mathcal{F}_r\right]\\
&\quad+2\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\left(\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))-\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right)\right\|^2\mid\mathcal{F}_r\right].
\end{aligned}
$$

Apply Lemma 9 to the first term on the right hand side:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2\mid\mathcal{F}_r\right]\\
&\le\frac{(n-B)}{B(n-1)}\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2\mid\mathcal{F}_r\right]\\
&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}(1-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\mid\mathcal{F}_r\right].
\end{aligned}
$$

Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{BH}\sum_{i\in\mathcal{B}_r}\sum_{h=1}^{H}(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\right\|^2\right]\\
&\leq \frac{2(n-B)}{B(n-1)}\frac{1}{nH}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathbb{E}\left[\left\|(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2\mid\mathcal{F}_r\right]+4\left\|\nabla F(\mathbf{w}_r)\right\|^2\\
&\quad+4E\left[\left\|\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\left(\nabla F_i(\mathbf{w}_{r,h}^i)-\nabla F_i(\mathbf{w}_r)\right)\right\|^2\mid\mathcal{F}_r\right]\\
&\quad+\frac{2(1+\alpha L)^2L^2}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h}^i)-\mathbf{u}_{r,h}^i\right\|^2\mid\mathcal{F}_r\right].
\end{aligned}$$

Based on Lemma 11, we have

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\left(I-\alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right)\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right\|^2\right]=\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{r,h}^i)\right\|^2\right]\\
&\leq\frac{2}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_r)\right\|^2\right]+\frac{2L_F^2}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2\right]\\
&\leq 2\gamma_F^2+2\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right]+\frac{2L_F^2}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2\right],
\end{aligned}$$

Put them together and use the tower property of expectation on both sides.

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{w}_{r+1})\right]\\
&\leq\mathbb{E}\left[F(\mathbf{w}_r)\right]-\frac{\tilde{\eta}}{2}\left(1-8\tilde{\eta}-\frac{8\tilde{\eta}(n-B)}{B(n-1)}\right)\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right]\\
&\quad+\frac{\tilde{\eta}}{2}\left(C_{\rho,G,L}+8\tilde{\eta}L_F^2+\frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2\right]\\
&\quad+8\tilde{\eta}\left(1+L^2\tilde{\eta}\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^i-\mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2\right]+\frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2+\frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}.
\end{aligned}$$

■

Lemma 20 *If $\eta\leq\frac{1}{2HL_F}$, it is satisfied that:*

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^i-\mathbf{w}_r\right\|^2\right]\\
&\leq 16\eta^2H(H-1)\left(\hat{\sigma}^2+2\gamma_F^2\right)+32\eta^2H^2L^2\left(\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^i-\mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2\right]\right)\\
&\quad+32\eta^2H(H-1)\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right].
\end{aligned}\tag{27}$$

Proof For the stochastic estimator $\widehat{\Delta}_{r,h}^i \coloneqq (I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2^i}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3^i}\mathcal{L}_i(\mathbf{u}_{r,h}^i)$ of the meta-gradient and $\Delta_{r,h}^i \coloneqq (I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)$, we have

$$\mathbb{E}\left[\widehat{\Delta}_{r,h}^i - \Delta_{r,h}^i\right] = \mathbb{E}\left[\mathbb{E}_{\mathcal{S}_2^i,\mathcal{S}_3^i}\left[\widehat{\Delta}_{r,h}^i - \Delta_{r,h}^i\right]\right] = 0.$$

Then, the local drift at iteration $h+1$, round r can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}_{r,h+1}^i - \mathbf{w}_r\|^2\right] \\
&\leq \left(1 + \frac{1}{H}\right)\mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] + \eta^2(1+H)\mathbb{E}\left[\|(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3}\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right] \\
&= \left(1 + \frac{1}{H}\right)\mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] + \eta^2(1+H)\mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right] \\
&\quad + \eta^2(1+H)\mathbb{E}\left[\|(I - \alpha\widehat{\nabla}^2_{\mathcal{S}_2}\mathcal{L}_i(\mathbf{w}_{r,h}^i))\widehat{\nabla}_{\mathcal{S}_3}\mathcal{L}_i(\mathbf{u}_{r,h}^i) - (I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right] \\
&\leq \left(1 + \frac{1}{H}\right)\mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] + 2\eta^2 H\mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right] + 2\eta^2 H\hat{\sigma}^2.
\end{aligned}$$

Next, we upper bound $\mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right]$. Note that we define $\mathbf{v}_i(\mathbf{w}_{r,h}^i) \coloneqq \mathbf{w}_{r,h}^i - \alpha\nabla\mathcal{L}_i(\mathbf{w}_{r,h}^i)$ and $\nabla F_i(\mathbf{w}_{r,h}^i) \coloneqq \left(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i)\right)\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))$.

$$\begin{aligned}
&\mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i)\|^2\right] \\
&\leq \mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\left(\nabla\mathcal{L}_i(\mathbf{u}_{r,h}^i) - \nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\right)\|^2\right] \\
&\qquad + \mathbb{E}\left[\|(I - \alpha\nabla^2\mathcal{L}_i(\mathbf{w}_{r,h}^i))\nabla\mathcal{L}_i(\mathbf{v}_i(\mathbf{w}_{r,h}^i))\|^2\right] \\
&\leq (1+\alpha L)^2 L^2\mathbb{E}\left[\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\|^2\right] + \mathbb{E}\left[\|\nabla F_i(\mathbf{w}_{r,h}^i)\|^2\right] \\
&\leq (1+\alpha L)^2 L^2\mathbb{E}\left[\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\|^2\right] + 2L_F^2\mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] + 2\mathbb{E}\left[\|\nabla F_i(\mathbf{w}_r)\|^2\right].
\end{aligned}$$

Thus, we have the following equation if $\eta \leq \frac{1}{2HL_F}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{w}_{r,h+1}^i - \mathbf{w}_r\right\|^2\right] \\
&\leq \left(1 + \frac{2}{H}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] + 2\eta^2 H\hat{\sigma}^2 \\
&\qquad + 2\eta^2 H\left(\frac{(1+\alpha L)^2 L^2}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\|^2\right] + \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla F_i(\mathbf{w}_r)\|^2\right]\right).
\end{aligned}$$

Based on Lemma 11, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla F_i(\mathbf{w}_r)\|^2\right] &= \mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla F_i(\mathbf{w}_r) - \nabla F(\mathbf{w}_r)\|^2\right] \\
&\leq \mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right] + \gamma_F^2.
\end{aligned}$$

Given that $\mathbf{w}_{r,1}^i = \mathbf{w}_r$ and $\alpha \leq 1/L$, and $H \geq 2$, we can obtain that

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\|\mathbf{w}_{r,h}^i - \mathbf{w}_r\|^2\right] \\
&\leq 16\eta^2 H(H-1)\left(\hat{\sigma}^2 + 2\gamma_F^2\right) + 32\eta^2 H^2 L^2\left(\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\|^2\right]\right) \\
&\qquad + 32\eta^2 H(H-1)\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right].
\end{aligned}$$

■

Lemma 21 *If $\alpha \leq 1/L$ and $\eta \leq \min\left\{\frac{1}{4HL_F}, \frac{\beta}{16L\mathbb{I}[\beta\in(0,1)]}\right\}$, we have*

$$\begin{aligned}
&\frac{1}{nHR}\sum_{i=1}^{n}\sum_{r=1}^{R}\sum_{h=1}^{H}\mathbb{E}\left[\|\mathbf{u}_{r,h}^i - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\|^2\right] \\
&\leq \frac{2\sigma_G^2}{\beta\left(\mathbb{I}[B<n]HS_0 + \mathbb{I}[B=n]HR\right)}\mathbb{I}[\beta\in(0,1)] + \frac{96\eta^2(\hat{\sigma}^2 + 2\gamma_F^2)}{\beta^2}\mathbb{I}[\beta\in(0,1)] + \frac{2\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&\qquad + \frac{192\eta^2}{\beta^2 R}\sum_{r=1}^{R}\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]\mathbb{I}[\beta\in(0,1)]. \qquad (28)
\end{aligned}$$

Proof We define that $\widehat{\mathbf{v}}_{r,h}^i \coloneqq \mathbf{w}_{r,h}^i - \alpha\widehat{\nabla}_{\mathcal{S}_1^i}\mathcal{L}_i(\mathbf{w}_{r,h})$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_{r,h+1}^i - \mathbf{v}_i(\mathbf{w}_{r,h+1}^i)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - (1-\beta)\mathbf{u}_{r,h}^i - \beta\widehat{\mathbf{v}}_{r,h+1}^i\right\|^2\right] \\
&= \mathbb{E}\left[\left\|(1-\beta)\left(\mathbf{v}_i(\mathbf{w}_{r,h}^i) - \mathbf{u}_{r,h}^i + \mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\right) + \beta(\mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - \widehat{\mathbf{v}}_{r,h+1}^i)\right\|^2\right]
\end{aligned}$$

Since $\mathbb{E}\left[\mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - \widehat{\mathbf{v}}_{r,h+1}^i\right] = 0$ and $\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - \widehat{\mathbf{v}}_{r,h+1}^i\right\|^2\right] \leq \frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_{r,h+1}^i - \mathbf{v}_i(\mathbf{w}_{r,h+1}^i)\right\|^2\right] \\
&\leq (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h}^i) - \mathbf{u}_{r,h}^i + \mathbf{v}_i(\mathbf{w}_{r,h+1}^i) - \mathbf{v}_i(\mathbf{w}_{r,h}^i)\right\|^2\right] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}. \\
&\leq (1-\beta)^2(1+\beta)\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h}^i) - \mathbf{u}_{r,h}^i\right\|^2\right] \\
&\qquad + 4(1-\beta)^2(1+1/\beta)(1+\alpha L)^2\mathbb{E}\left[\left\|\mathbf{w}_{r,h+1}^i - \mathbf{w}_{r,h}^i\right\|^2\right]\mathbb{I}[\beta\in(0,1)] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&\leq (1-\beta)\mathbb{E}\left[\left\|\mathbf{v}_i(\mathbf{w}_{r,h}^i) - \mathbf{u}_{r,h}^i\right\|^2\right] + \frac{4\eta^2(1+\alpha L)^2}{\beta}\mathbb{E}\left[\left\|\widehat{\Delta}_{r,h}^i\right\|^2\right]\mathbb{I}[\beta\in(0,1)] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}.
\end{aligned}$$

As done in the proof of Lemma 20, the term $\mathbb{E}\left[\left\|\widehat{\Delta}^i_{r,h}\right\|^2\right]$ can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\Delta}^i_{r,h}\|^2\right] \\
&\leq \hat{\sigma}^2 + (1+\alpha L)^2 L^2 \mathbb{E}\left[\|\mathbf{u}^i_{r,h} - \mathbf{v}_i(\mathbf{w}^i_{r,h})\|^2\right] + 2L_F^2 \mathbb{E}\left[\|\mathbf{w}^i_{r,h} - \mathbf{w}_r\|^2\right] + 2\mathbb{E}\left[\|\nabla F_i(\mathbf{w}_r)\|^2\right].
\end{aligned}$$

If $\alpha \leq 1/L$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}^i_{r,h+1} - \mathbf{v}_i(\mathbf{w}^i_{r,h+1})\|^2\right] \\
&\leq \left(1 - \beta + \frac{64\eta^2 L^2}{\beta}\mathbb{I}[\beta \in (0,1)]\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}^i_{r,h} - \mathbf{v}_i(\mathbf{w}^i_{r,h})\|^2\right] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} + \frac{16\eta^2}{\beta}\hat{\sigma}^2\mathbb{I}[\beta \in (0,1)] \\
&\quad + \frac{32L_F^2\eta^2}{\beta}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{w}^i_{r,h} - \mathbf{w}_r\|^2\right]\mathbb{I}[\beta \in (0,1)] + \frac{32\eta^2}{\beta}\left(\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right] + \gamma_F^2\right)\mathbb{I}[\beta \in (0,1)].
\end{aligned}$$

Telescope it from $h = 1$ to H and divide both sides by H.

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{u}^i_{r,h+1} - \mathbf{v}_i(\mathbf{w}^i_{r,h+1})\|^2\right] \\
&\leq \left(1 - \beta + \frac{64\eta^2 L^2}{\beta}\mathbb{I}[\beta \in (0,1)]\right)\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{u}^i_{r,h} - \mathbf{v}_i(\mathbf{w}^i_{r,h})\|^2\right] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} + \frac{16\eta^2}{\beta}\hat{\sigma}^2\mathbb{I}[\beta \in (0,1)] \\
&\quad + \frac{32L_F^2\eta^2}{\beta}\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{w}^i_{r,h} - \mathbf{w}_r\|^2\right]\mathbb{I}[\beta \in (0,1)] + \frac{32\eta^2}{\beta}\left(\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right] + \gamma_F^2\right)\mathbb{I}[\beta \in (0,1)].
\end{aligned}$$

Applying Lemma 20 and setting $\eta \leq \min\left\{\frac{1}{4HL_F}, \frac{\beta}{16L\mathbb{I}[\beta\in(0,1)]}\right\}$ leads to

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{u}^i_{r,h+1} - \mathbf{v}_i(\mathbf{w}^i_{r,h+1})\|^2\right] \\
&\leq \left(1 - \beta + \frac{64\eta^2 L^2}{\beta}(1 + 16\eta^2H^2L_F^2)\mathbb{I}[\beta \in (0,1)]\right)\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{u}^i_{r,h} - \mathbf{v}_i(\mathbf{w}^i_{r,h})\|^2\right] \\
&\qquad + \frac{16\eta^2}{\beta}(1 + 32L_F^2\eta^2H(H-1))\left(\hat{\sigma}^2 + 2\gamma_F^2\right)\mathbb{I}[\beta \in (0,1)] + \beta^2\frac{\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&\qquad + \frac{32\eta^2}{\beta}\left(1 + 32L_F^2\eta^2H(H-1)\right)\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]\mathbb{I}[\beta \in (0,1)] \\
&\leq \left(1 - \frac{\beta}{2}\right)\frac{1}{nH}\sum_{i=1}^n\sum_{h=1}^H \mathbb{E}\left[\|\mathbf{u}^i_{r,h} - \mathbf{v}_i(\mathbf{w}^i_{r,h})\|^2\right]\mathbb{I}[\beta \in (0,1)] \\
&\qquad + \frac{48\eta^2\left(\hat{\sigma}^2 + 2\gamma_F^2\right)}{\beta}\mathbb{I}[\beta \in (0,1)] + \frac{\beta^2\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} + \frac{96\eta^2}{\beta}\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]\mathbb{I}[\beta \in (0,1)].
\end{aligned}$$

Re-arrange the terms, telescope it from $r = 1$ to R, and divide both sides by $\beta R/2$

$$\frac{1}{nHR}\sum_{i=1}^{n}\sum_{r=1}^{R}\sum_{h=1}^{H}\mathbb{E}\left[\|\mathbf{u}_{r,h}^{i}-\mathbf{v}_i(\mathbf{w}_{r,h}^{i})\|^2\right] \le \frac{2\sigma_G^2}{\beta\left(\mathbb{I}[B<n]HS_0+\mathbb{I}[B=n]HR\right)}\mathbb{I}[\beta\in(0,1)]$$
$$+\frac{96\eta^2(\hat{\sigma}^2+2\gamma_F^2)}{\beta^2}\mathbb{I}[\beta\in(0,1)]+\frac{2\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|}+\frac{192\eta^2}{\beta^2 R}\sum_{r=1}^{R}\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]\mathbb{I}[\beta\in(0,1)].$$

■

Proof [Proof of Theorem 7] Based on (24) and (27), we have

$$\mathbb{E}\left[F(\mathbf{w}_{r+1})\right] \le \mathbb{E}\left[F(\mathbf{w}_r)\right]-\frac{\tilde{\eta}}{2}\left(1-8\tilde{\eta}-\frac{8\tilde{\eta}(n-B)}{B(n-1)}\right)\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]$$
$$+\frac{\tilde{\eta}}{2}\left(C_{\rho,G,L}+8\tilde{\eta}L_F^2+\frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{w}_{r,h}^{i}-\mathbf{w}_r\right\|^2\right]$$
$$+8\tilde{\eta}(1+L^2\tilde{\eta})\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^{i}-\mathbf{v}_i(\mathbf{w}_{r,h}^{i})\right\|^2\right]+\frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2+\frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}$$
$$\le \mathbb{E}\left[F(\mathbf{w}_r)\right]+8\tilde{\eta}\eta^2H(H-1)\left(C_{\rho,G,L}+8\tilde{\eta}L_F^2+\frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)(\hat{\sigma}^2+2\gamma_F^2)+\frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2$$
$$-\frac{\tilde{\eta}}{2}\left(1-8\tilde{\eta}-\frac{8\tilde{\eta}(n-B)}{B(n-1)}-32\tilde{\eta}^2\left(C_{\rho,G,L}+8\tilde{\eta}L_F^2+\frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)\right)\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]+\frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}$$
$$+8\tilde{\eta}\left(1+L^2\tilde{\eta}+2\tilde{\eta}^2L^2\left(C_{\rho,G,L}+8\tilde{\eta}L_F^2+\frac{8\tilde{\eta}(n-B)L_F^2}{B(n-1)}\right)\right)\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^{i}-\mathbf{v}_i(\mathbf{w}_{r,h}^{i})\right\|^2\right]$$
$$\le \mathbb{E}\left[F(\mathbf{w}_r)\right]+8\tilde{\eta}\eta^2H(H-1)C_1(\hat{\sigma}^2+2\gamma_F^2)+\frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2+\frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}-\frac{\tilde{\eta}}{2}\left(1-8C_2\tilde{\eta}\right)\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]$$
$$+8\tilde{\eta}C_3\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^{i}-\mathbf{v}_i(\mathbf{w}_{r,h}^{i})\right\|^2\right].$$

where we set $\eta \le \frac{1}{4HL_F}$ and define $C_1 \coloneqq C_{\rho,G,L}+2L_F+\frac{2(n-B)L_F}{B(n-1)}$, $C_2 \coloneqq 1+\frac{n-B}{B(n-1)}+\frac{C_1}{4L_F^2}$, and $C_3 \coloneqq 1+\frac{L^2}{4L_F}+\frac{L^2C_1}{8L_F^2}$. Telescoping the equation above from round $r = 1$ to R and dividing both sides by R leads to

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[F(\mathbf{w}_{r+1})\right] \le \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[F(\mathbf{w}_r)\right]+8\tilde{\eta}\eta^2H(H-1)C_1(\hat{\sigma}^2+2\gamma_F^2)+\frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2$$
$$+\frac{\tilde{\eta}^2\hat{\sigma}^2}{BH}-\frac{\tilde{\eta}}{2}\left(1-8C_2\tilde{\eta}\right)\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\|\nabla F(\mathbf{w}_r)\|^2\right]$$
$$+8\tilde{\eta}C_3\frac{1}{nHR}\sum_{i=1}^{n}\sum_{r=1}^{R}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\mathbf{u}_{r,h}^{i}-\mathbf{v}_i(\mathbf{w}_{r,h}^{i})\right\|^2\right].$$

The last term above can be upper bounded by Lemma 21.

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[F(\mathbf{w}_{r+1})\right] \\
&\leq \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[F(\mathbf{w}_r)\right] + 8\tilde{\eta}\eta^2 H(H-1)C_1(\hat{\sigma}^2+2\gamma_F^2) + \frac{4(n-B)}{B(n-1)}\tilde{\eta}^2\gamma_F^2 \\
&+ \frac{\tilde{\eta}^2\hat{\sigma}^2}{BH} - \frac{\tilde{\eta}}{2}\left(1 - 8C_2\tilde{\eta} - \frac{6144\eta^2 C_3}{\beta^2}\mathbb{I}[\beta\in(0,1)]\right)\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right] + \frac{16\tilde{\eta}C_3\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&+ \frac{16\tilde{\eta}C_3\sigma_G^2}{\beta\left(\mathbb{I}[B<n]HS_0 + \mathbb{I}[B=n]HR\right)}\mathbb{I}[\beta\in(0,1)] + \frac{768\tilde{\eta}C_3\eta^2(\hat{\sigma}^2+2\gamma_F^2)}{\beta^2}\mathbb{I}[\beta\in(0,1)].
\end{aligned}$$

If we set $\tilde{\eta} = \eta H \leq \frac{1}{32C_2}$ and $\eta \leq \frac{\beta}{\sqrt{24576C_3\mathbb{I}[\beta\in(0,1)]}}$, we have

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_r)\right\|^2\right] \\
&\leq \frac{4F(\mathbf{w}_1)}{\eta T} + 32\eta^2 H(H-1)C_1(\hat{\sigma}^2+2\gamma_F^2) + \frac{4\eta}{B}\left(\hat{\sigma}^2 + \frac{4(n-B)}{(n-1)}H\gamma_F^2\right) + \frac{64C_3\beta\alpha^2\sigma_G^2}{|\mathcal{S}_1^i|} \\
&+ \frac{64C_3\sigma_G^2}{\beta\left(\mathbb{I}[B<n]HS_0 + \mathbb{I}[B=n]HR\right)}\mathbb{I}[\beta\in(0,1)] + \frac{3072C_3\eta^2(\hat{\sigma}^2+2\gamma_F^2)}{\beta^2}\mathbb{I}[\beta\in(0,1)],
\end{aligned}$$

where we use the fact $\tilde{\eta}R = \eta HR = \eta T$. Then, we can define $C_4 \coloneqq \frac{1}{4\max\{L_F, 8C_2\}}$, $C_5 \coloneqq \frac{1}{16\max\{L,\sqrt{96C_3}\}}$, and set $\eta \leq \min\left\{\frac{C_4}{H}, \frac{C_5\beta}{\mathbb{I}[\beta\in(0,1)]}\right\}$. ■

Appendix E. Additional Theoretical Results

In this section, we provide some other theoretical results that we obtained.

E.1 MAML, BSGD, and BSpiderBoost in the finite #tasks case

In original papers of MAML (Fallah et al., 2020a) and BSGD/BSpiderBoost (Hu et al., 2020), the convergence rates are only available for the case that the number of tasks n is infinite (e.g., the tasks are online). However, it is easy to extend their results to the finite n case. For example, Equation (97) of Fallah et al. (2020a) uses the fact

$$\mathbb{E}\left[\frac{1}{B}\sum_{i\in\mathcal{B}}X_{\xi_i}\right] = X, \quad \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}}X_{\xi_i}\right\|^2\right] \leq \frac{1}{B}\mathbb{E}_{\xi}\left[\left\|X_{\xi}\right\|^2\right] + \left\|X\right\|^2,$$

for $\mathbb{E}\left[X_{\xi_i}\right] = X$. This further leads to requirement on batch size $B = \mathcal{O}(1/\epsilon^2)$ in the final result to ensure the convergence. In the finite n case, we instead use the finite n counterpart in Lemma 9. Then, we can conclude that MAML needs $B = n$ to ensure convergence if we follow the rest part of the analysis of Fallah et al. (2020a).

E.2 Comparison of convergence rates in the infinite n case

As demonstrated in the main paper, our MOML only works when the number of tasks is finite. However, our LocalMOML works for both finite and infinite n if it is implemented on a single machine. In Table 2, we compare it with existing results.

E.3 Problematic proof in Fallah et al. (2020b) of Per-FedAvg

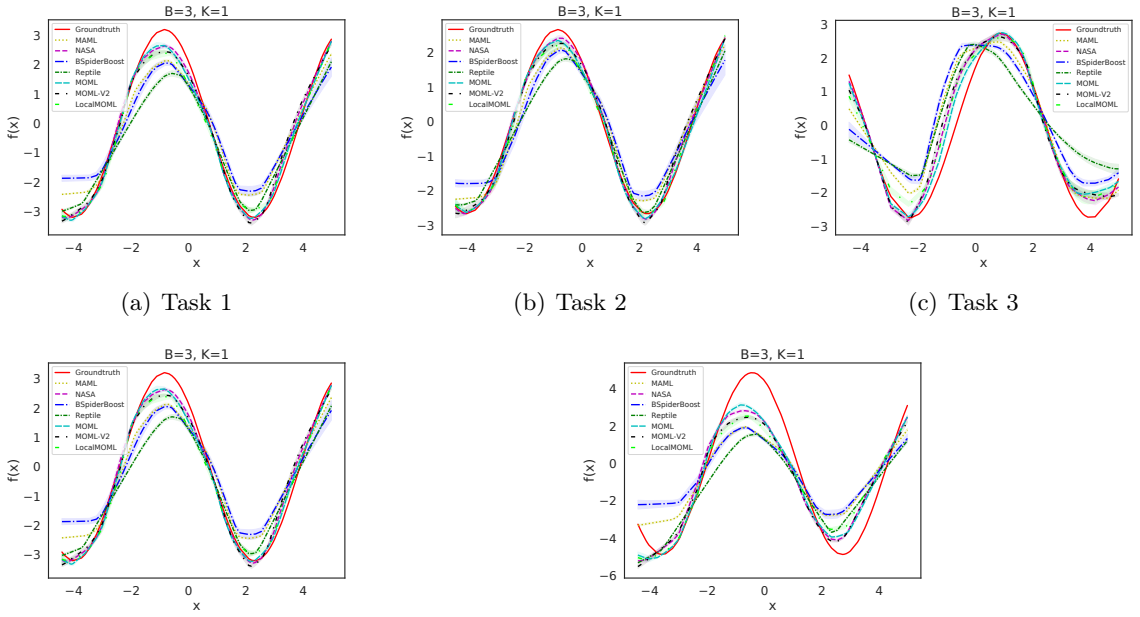
The convergence analysis in Fallah et al. (2020b) is problematic starting from Equation (99) in their paper (We follow their notations below):

$$\mathbb{E}\left[\left(\frac{1}{\tau n}\sum_{i\in\mathcal{A}_k}\nabla F_i(\bar{w}_{k+1,t})\right) \mid \mathcal{F}_{k+1}^t\right] = \frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\bar{w}_{k+1,t}).$$

The equality above is wrong unless $|\mathcal{A}_k| = n$ (full client participation) because $\bar{w}_{k+1,t} = \frac{1}{\tau n}\sum_{i\in\mathcal{A}_k} w_{k+1,t}^i$ also depends on the randomly sampled batch of clients $\mathcal{A}_k$. This fault makes their proof cannot proceed. In this paper, we corrected this issue.

Appendix F. Additional Experimental Results of Sinewave Regression

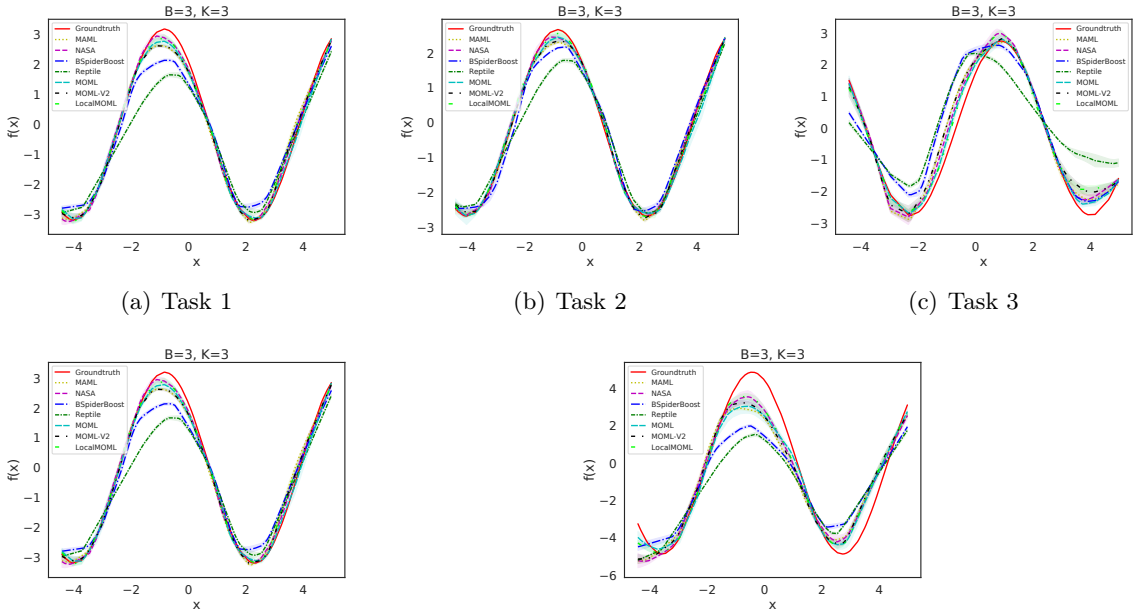
We also provide the fitted sinusoid curves on unseen tasks. The curves when $K = 1$ on 5 unseen tasks can be found in Figure 2 and those of $K = 3$ can be found in Figure 3.



(a) Task 1 (b) Task 2 (c) Task 3

(d) Task 4 (e) Task 5

Figure 2: Fitted Sinusoid Curves on five unseen tasks when $K = 1$.



(a) Task 1 (b) Task 2 (c) Task 3

(d) Task 4 (e) Task 5

Figure 3: Fitted Sinusoid Curves on five unseen tasks when $K = 5$.