First-order Personalized Federated Meta-Learning via Over-the-Air Computations

Zeshen Li[1], Zihan Chen[2], Howard H. Yang[1]*

[1]Zhejiang University, China
[2]Singapore University of Technology and Design
zeshen.22@intl.zju.edu.cn, zihan_chen@alumni.sutd.edu.sg, haoyang@intl.zju.edu.cn

Abstract

Federated learning (FL) is an emerging approach in machine learning that enables large-scale distributed model training without sharing local private data. However, training a generic global model often fails to meet the personalized demands of all clients over heterogeneous networks. Gradient-based meta-learning, especially MAML, has become a viable solution for this objective. One issue with MAML is the computational and memory burden introduced by the second-order information needed to compute the meta-gradient. Additionally, frequent communication between clients and server for model update in FL systems presents a significant communication bottleneck in wireless networks. In this paper, we propose a novel personalized federated meta-learning system that leverages only first-order information and utilizes over-the-air computations to improve communication efficiency. We prove the convergence of our algorithm under non-convex conditions and demonstrate its effectiveness through extensive numerical experiments.

Introduction

Conventional federated learning (FL) aims to train a shared global model across all clients (Fallah, Mokhtari, and Ozdaglar 2020). However, due to the heterogeneous nature of networks, a single global model may fail to obtain desirable performance for individual clients (Tan et al. 2022). Specifically, when the local datasets of the users are non-i.i.d. (where i.i.d. stands for independent identically distributed), the model tends to favor some of the users while heavily degrading the performance of others (Li et al. 2019). To address this issue, it is necessary to improve the performance of FL models on clients with heterogeneous data through personalized approaches (Zhang et al. 2023), a concept known as personalized federated learning (PFL).

Furthermore, for modern artificial intelligence applications, the training paradigm has recently shifted to pre-training followed by fine-tuning (Wen, Xing, and Simeone 2024). Motivated by this principle, by performing fine-tuning of a unified pre-trained model on each local client, we can naturally enhance the performance of customized FL models on diverse client data. Specifically, the goal of the FL system shifts to training a pre-trained model, where each client performs its unique local fine-tuning process by capturing the local-global model relationship and cross-client knowledge, which can be regarded as a multi-task learning framework (Smith et al. 2017).

A general framework in which pre-training and fine-tuning can be formalized is meta-learning. In meta-learning, data from different tasks are used to pre-train a model with the aim of ensuring that the pre-trained model can be efficiently fine-tuned based on limited data for a new task (Chen et al. 2023a). Model Agnostic Meta Learning (MAML) introduced by (Finn, Abbeel, and Levine 2017) is a gradient-based meta learning algorithm, which runs in two connected stages: meta-training and meta-testing. Meta-training uses gradient descent to learn an initial model that can quickly adapt to a range of possible tasks. Meta-testing then involves training this initial model on a specific task to evaluate its performance. For FL systems, we can approximate heterogeneous clients as tasks in MAML (Fallah, Mokhtari, and Ozdaglar 2020). The global model training process in FL can be seen as meta-training in MAML, while the client-specific personalization based on the global FL model can be understood as meta-testing (Jiang et al. 2019).

Based on the MAML algorithm, we can restructure the FL system by shifting its objective from finding a single model that is optimal on average for all clients to finding an optimal initial pre-trained model. However, the original MAML algorithm requires calculating second-order gradients, specifically the Hessian matrix, in each update (Chayti and Jaggi 2024). This typically consumes substantial computational resources, which is especially challenging for resource-constrained FL systems in wireless networks.

Communication efficiency serves as another issue hindering the scalability of the deployment of FL systems over wireless networks (Yang et al. Dec. 2021,M). A typical training process for a generic global model to converge requires hundreds or even thousands of communication rounds among the massively distributed clients and the edge server, where the iterative exchange of model parameters incurs substantial communication overhead (McMahan et al. Apr. 2017; Li et al. May. 2020). One way to address this issue is to integrate analog over-the-air (OTA) computations into the FL model training, exploiting the superposition properties of analog transmissions, so as to achieve automatic "one-shot"

*Corresponding author: Howard H. Yang.

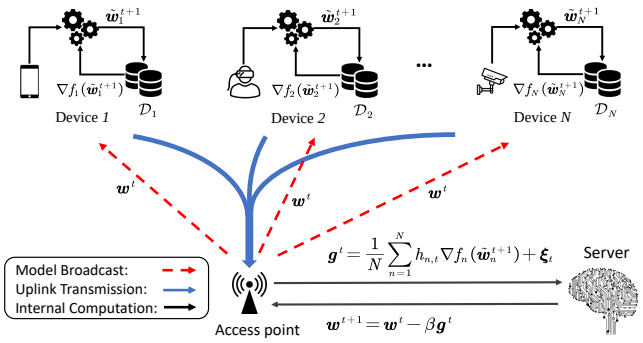

Figure 1: An overview of gradient descent-based first-order personalized federated meta-learning system, which do not require the use of the Hessian matrix.

aggregation for model updates (Guo et al. Dec. 2021; Chen, Yang, and Quek 2023). However, the random channel fading and interference would concurrently be induced into the aggregated gradients or model updates, leading to performance degradation such as the slower convergence rate and instability (Sery and Cohen Apr. 2020; Chen et al. 2023b).

In view of the above challenges and considerations, we propose a first-order federated meta-learning framework in the context of analog transmissions. Unlike other OTA FL frameworks (Wen, Xing, and Simeone 2024), we adopt the FOMAML (First-Order MAML) approach (Chen et al. 2018), which does not require second-order information. This significantly reduces computational complexity during the training phase with minimal impact on the model's final performance. We also provide a convergence analysis of our algorithm under non-convex conditions. Both theoretical and numerical results validate the gain of our design.

System model

We consider a federated edge learning system consisting of an edge server and N clients. The clients communicate with the edge server through wireless channels. Each client n, $n \in \{1, \cdots, N\}$, holds a local dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m_n}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ represent the input data sample and the corresponding response, respectively.

The goal of each device n is to minimize the emprical loss function constructed by the local dataset of client n, given by

$$f_n(\boldsymbol{w}) = \frac{1}{m_n} \sum_{i=1}^{m_n} \ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) \tag{1}$$

in which $\ell(\cdot)$ represents the loss evaluated at one pair of input data samples. In a conventional FL system, the edge server needs to coordinate clients to jointly optimize the following objective function:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad f(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} f_n(\boldsymbol{w}). \tag{2}$$

However, in contrast to the traditional setting, the goal in meta-learning is not finding a model which performs well on all the tasks in expectation. Instead, we assume that we have a limited computational budget to update our model after a new task arrives, and in this new setting, we look for an initialization which performs well after it is updated with respect to this new task, possibly by one or a few steps of gradient descent (Finn, Abbeel, and Levine 2017). In this paper, we assume the use of a single-step gradient descent to simplify the analysis. Then problem (2) changes to

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad F(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} f_n(\boldsymbol{w} - \alpha \nabla f_n(\boldsymbol{w})) \tag{3}$$

where $\alpha \geq 0$ is the inner stepsize. By solving (3), we can find an initial model (meta-model) which is trained in a way that after one step of local gradient leads to a good personalized model for each individual clients. In the next section, we will describe how to solve the above optimization problem in wireless networks.

Model training procedure

This section details the meta model training process based on OTA computing schemes. First, note that (3) can be rewritten as the average of meta-functions $F_1, ..., F_n$, where the meta-function F_n associated with n is defined as

$$F_n(\boldsymbol{w}) = f_n(\boldsymbol{w} - \alpha \nabla f_n(\boldsymbol{w})). \tag{4}$$

Then, the meta-gradient $\nabla F_n(\boldsymbol{w})$ is given by

$$\nabla F_n(\boldsymbol{w}) = \left(I - \alpha \nabla^2 f_n(\boldsymbol{w})\right) \nabla f_n(\boldsymbol{w} - \alpha \nabla f_n(\boldsymbol{w})). \tag{5}$$

Using meta-gradients, we can solve optimization problem (2) through gradient descent method. However, computing the Hessian matrix in each iteration is usually computationally expensive. For example, the computational complexity of calculating the gradient $\nabla f_n(\boldsymbol{w})$ can be considered $O(d)$, while the computational complexity of calculating the second-order gradient $\nabla^2 f_n(\boldsymbol{w})$ is typically $O(d^2)$ (Nocedal and Wright 1999). For high-dimensional machine learning models, computing the Hessian matrix undoubtedly increases the computational complexity, thereby prolonging the training time. Therefore, it is meaningful to adopt a first-order method that does not significantly reduce the final training accuracy to avoid this issue.

In this paper, inspired by FOMAML (Chen et al. 2018), we replace the meta-gradient $\nabla F_n(\boldsymbol{w})$ with a first-order approximation $\nabla \hat{F}_n(\boldsymbol{w})$ that ignores the Hessian term

$$\nabla \hat{F}_n(\boldsymbol{w}) = \nabla f_n(\boldsymbol{w} - \alpha \nabla f_n(\boldsymbol{w})). \tag{6}$$

We will use this first-order approximation instead of the actual meta-gradient in the subsequent training process. In the following sections, we will demonstrate through theoretical analysis and experiments that this first-order algorithm can also ensure convergence, while significantly improving the algorithm's running speed without notably decreasing the final accuracy. The detailed training procedure is as follows:

1) Local Model Training: Without loss of generality, we assume the system has progressed to the t-th round of global

training, where the clients just received the global model parameters $\boldsymbol{w}^t$ from the edge server. Then, each client n performs two step of gradient descent

$$\tilde{\boldsymbol{w}}_n^{t+1} = \boldsymbol{w}^t - \alpha \nabla f_n\left(\boldsymbol{w}^t\right). \tag{7}$$

$$\boldsymbol{w}_n^{t+1} = \boldsymbol{w}^t - \beta \nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) \tag{8}$$

where $\beta \geq 0$ is the outer stepsize and $\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)$ will be uploaded to the edge server via analog transmissions.

2) Analog Gradient Aggregation: We consider the clients adopt analog transmissions to upload their locally trained parameters. Specifically, once $\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)$ is computed, client n modulates it entry-by-entry onto the magnitudes of a common set of orthogonal baseband waveforms, forming the following analog signal

$$x_n(s) = \left\langle \boldsymbol{c}(s), \nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) \right\rangle \tag{9}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors and $\boldsymbol{c}(s) = (c_1(s), ..., c_d(s)), s \in [0, \tau]$ has its entries satisfying

$$\int_0^\tau c_i^2(s) ds = 1, \;\; i = 1, 2, ..., d \tag{10}$$

$$\int_0^\tau c_i(s) c_j(s) ds = 0, \;\; i \neq j. \tag{11}$$

Then, the clients transmit their analog signals concurrently to the edge server.

Due to the superposition property of electromagnetic waves, the signal received by the edge server takes the following form

$$y(s) = \sum_{n=1}^{N} h_{n,t} x_n(s) + \xi(s), \tag{12}$$

where $h_{n,t}$ is the channel fading experienced by client n and $\xi(s)$ denotes the additive noise. In this work, we assume the channel fading is i.i.d. across clients, with mean μ_h and variance σ_h^2. Besides, the transmit power of each client is set to compensate for the large-scale path loss. Additionally, we assume the noise follows a Gaussian distribution with variance σ_g^2.

This received signal will be passed through a bank of match filters, with each branch tuning to $c_i(s), i = 1, 2, .., d$. On the output side, the server obtains the following vector:

$$\boldsymbol{g}^t = \frac{1}{N} \sum_{n=1}^{N} h_{n,t} \nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t, \tag{13}$$

in which $\boldsymbol{\xi}_t$ is a d-dimensional random vector with each entry being i.i.d. and follows the Gaussian distribution.

3) Global Model Update: Using $\boldsymbol{g}^t$, the server updates the global model as follows:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \beta \boldsymbol{g}^t. \tag{14}$$

After this, the server broadcasts the $\boldsymbol{w}^{t+1}$ to all the clients for the next round of local computing. Such a process will iterate for T rounds until the model converges. We summarize the proposed framework in Algorithm 1.

Algorithm 1: First-order federated meta-learning with Over-the-Air Computations

Input: Initial global model $\boldsymbol{w}^0$, communication round T, inner stepsize α, outer stepsize β
Output: Global meta model $\boldsymbol{w}^T$

```
for t = 0, 1, 2 to T − 1 do
    for n = 1, 2, to N in parallel do
    # First-order local meta-model update
        w̃_n^{t+1} = w^t − α∇f_n(w^t)
        w_n^{t+1} = w^t − β∇f_n(w̃_n^{t+1})
    # Noisy aggregation via OTA computations
    g^t = (1/N) Σ_{n=1}^N h_{n,t} ∇f_n(w̃_n^{t+1}) + ξ_t
    # Global model update
    w^{t+1} = w^t − β g^t
return w^T
```

Convergence analysis

In this section, we analyze the convergence rate of the proposed model training framework, which quantifies the training efficiency.

To facilitate the analysis, we make the following assumptions.

Assumption 1. *For every $n \in \{1, \cdots, N\}$, f_n is L-Lipschitz, i.e., for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it is satisfied:*

$$\|\nabla f_n(\boldsymbol{x}) - \nabla f_n(\boldsymbol{y})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2, \tag{15}$$

where L is a non-negative constant.

Assumption 2. *The gradients of $f_n(\boldsymbol{w})$ are bounded; namely, there exists a constant G such that*

$$\|\nabla f(\boldsymbol{w})\|_2 \leq G, \quad \forall \boldsymbol{w} \in \mathbb{R}^d, \; n = 1, ..., N. \tag{16}$$

Assumption 3. *The dissimilarity of $f_n(\boldsymbol{w})$ and $f(\boldsymbol{w})$ is bounded as follows*

$$\|f_n(\boldsymbol{w}) - f(\boldsymbol{w})\|^2 \leq \sigma. \tag{17}$$

We are now in position to present the main theoretical finding of this paper.

Theorem 1. *Define F^* as the optimal value of optimization problem (3). Under the considered federated meta-learning system, the global parameters converge as follows:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \leq \frac{2}{\beta T}\left(\mathbb{E}\left[\nabla F(\boldsymbol{w}^0)\right] - F^*\right) + C_1 + \beta L C_2 \tag{18}$$

in which C_1,C_2 is given by

$$C_1 = \frac{8}{N}\left((\mu_h - 1)^2 + \alpha^2 L^2\right)\left(1 + \alpha^2 L^2\right)\left(G^2 + \sigma^2\right) \tag{19}$$

$$C_2 = \frac{4}{N}\left(\mu_h^2 + \sigma_h^2\right)\left(1 + \alpha^2 L^2\right)\left(G^2 + \sigma^2\right) + d\sigma_g \tag{20}$$

Proof. Please see Appendix. □

Remark 1. *Despite distortions from channel fading and interference, the global meta models converge under the proposed model training framework, with a convergence rate at the order of* $O\left(\frac{1}{T}\right)$.

Remark 2. *Greater heterogeneity among clients, represented by a larger* σ*, will lead to a larger convergence error for the global meta-model. Similarly, a larger internal learning rate* α *will also reduce convergence performance.*

Remark 3. *An increase in channel fading variance and additive noise variance is detrimental to the convergence rate. In contrast, a channel fading mean closer to 1 and an increase in the number of participating clients contribute to faster convergence.*

Numerical results

In this section, we evaluate the performance of our proposed framework. First, we introduce the experimental setup, followed by the discussion of the performance evaluation results.

Experiment Setup. We evaluate the performance of our framework by carrying out an image classification task: training a CNN on the EMNIST dataset (Cohen et al. May. 2017). The EMNIST balanced dataset contains 131,600 data samples collected from 47 categories. We divide the full dataset into two portions, each with 112,800 and 18,800 images, for training and test, respectively. Throughout the experiments, we set the number of clients to be $N = 50$. Accordingly, for the training set, we construct 50 sets of disjoint data samples, formed in a non-i.i.d. manner, and assign them to the clients. Specifically, the non-i.i.d. data partitions are implemented using a symmetric Dirichlet distribution, where the parameter Dir (Xu et al. 2022; Wang et al. 2024) controls the degree of data heterogeneity: the smaller the Dir, the higher degrees of non-i.i.d.ness in the data distribution. Unless otherwise stated, we set $Dir = 0.1$ throughout the experiments. For performance evaluation, following the approach of MAML, we perform a single step of gradient descent on the client's local training set after each global meta-learning model update. The model is then evaluated on the client's private local test set to assess its performance. Note that these local test sets are kept on each client with the same distribution as the training sets for model training. Each test set is non-i.i.d. relative to the others, simulating data heterogeneity in the FL system. Since there are multiple clients in the FL system, we use the average test accuracy of each client as the metric for local model performance evaluation. Throughout the experiments, unless otherwise specified, we set $\alpha = 0.01$, $\beta = 0.03$. Moreover, we employ Rayleigh fading to model the channel gain and use a Gaussian distribution to characterize the channel interference. All experiments are implemented using PyTorch on NVIDIA RTX 3090 GPU.

Performance evaluation. In Fig 2, we first compare the FOMAML algorithm based on analog OTA computations with the conventional digital communication-based algorithm. The purpose of this experiment is to verify the stability of the FOMAML algorithm under different communication modes. Clearly, the analog transmission method sig-

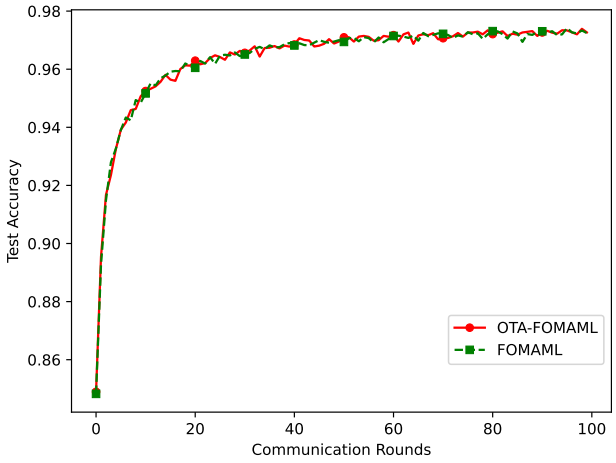

Figure 2: Test accuracy of FOMAML under analog over-the-air computations and conventional digital communication.

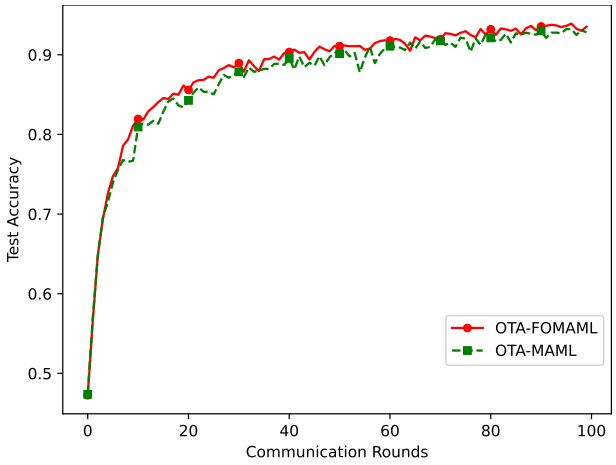

Figure 3: Comparison of test accuracy between first-order and second-order algorithms .

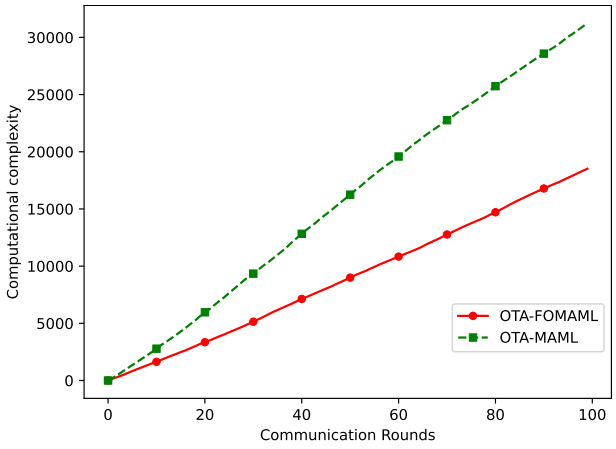

Figure 4: Comparison of computational complexity between first-order and second-order algorithms, where the complexity is represented by the algorithm's running time under the same settings.

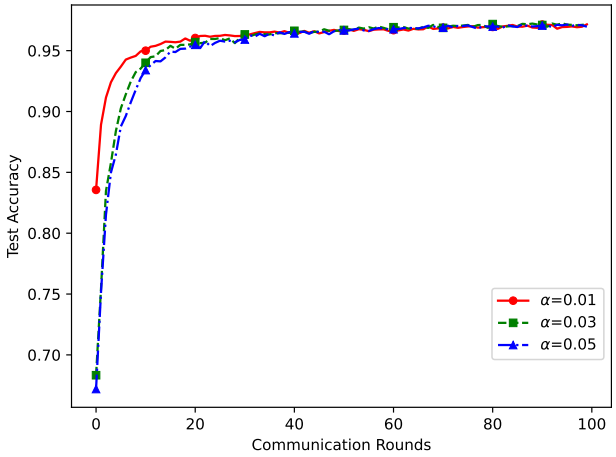

Figure 5: Performance of OTA-FOMAML under different inner stepsize α.

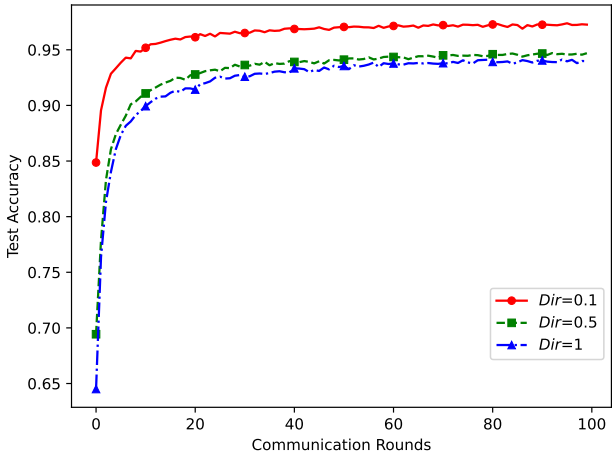

Figure 6: Performance of OTA-FOMAML under different Dir.

nificantly improves communication efficiency without noticeably impacting model accuracy. The channel fading and additive noise introduced by the OTA process did not significantly affect the final performance of the personalized federated model.

Subsequently, in Fig 3 and Fig 4, we compare the FOMAML algorithm with the MAML algorithm in the context of analog transmissions, which requires second-order information. We use the running time of the algorithm under the same settings to represent the computational resources consumed by the algorithm, i.e., the algorithm's complexity. As we can see, the FOMAML algorithm, which only requires first-order information, substantially reduces computational complexity while maintaining model performance comparable to MAML. As shown in Fig 4, our algorithm reduces the computational complexity by half compared with OTA-MAML algorithm. At the same time, our first-order algorithm consistently outperforms the second-order algorithm, which could be due to the noise introduced by the OTA process affecting the more sensitive second-order algorithm. A more detailed investigation of this issue will be left for future work.

Finally, we conduct experiments to test key parameters of our algorithm in Fig 5 and Fig 6. Consistent with theoretical analysis, a smaller α enhances model convergence performance. Additionally, greater data heterogeneity, indicating better consistency among clients' local datasets, improves the performance of personalized models after fine-tuning the meta-model.

Conclusion

In this paper, we proposed a personalized federated meta-learning system based on first-order information and OTA computations. This system leverages analog OTA computations to address communication efficiency challenges in FL systems and reduces the high computational demands of the conventional MAML algorithm by avoiding second-order information. Both theoretical analysis and experimental results demonstrate the effectiveness of this framework. To the best of our knowledge, this is the first use of a first-order meta-learning method in a personalized FL system based on OTA computations. Future work may explore other advanced first-order MAML methods to enhance robustness further.

References

Chayti, E. M.; and Jaggi, M. 2024. A New First-Order Meta-Learning Algorithm with Convergence Guarantees. *arXiv preprint arXiv:2409.03682*.

Chen, F.; Luo, M.; Dong, Z.; Li, Z.; and He, X. 2018. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*.

Chen, L.; Jose, S. T.; Nikoloska, I.; Park, S.; Chen, T.; Simeone, O.; et al. 2023a. Learning with limited samples: Meta-learning and applications to communication systems. *Foundations and Trends® in Signal Processing*, 17(2): 79–208.

Chen, Z.; Li, Z.; Yang, H. H.; and Quek, T. Q. 2023b. Personalizing federated learning with over-the-air computations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chen, Z.; Yang, H. H.; and Quek, T. Q. 2023. Edge intelligence over the air: Two faces of interference in federated learning. *IEEE Communications Magazine*.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. May. 2017. EMNIST: Extending MNIST to handwritten letters. In *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, 2921–2926. Anchorage, Alaska, USA.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33: 3557–3568.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Guo, H.; Zhu, Y.; Ma, H.; Lau, V. K.; Huang, K.; Li, X.; Nong, H.; and Zhou, M. Dec. 2021. Over-the-Air Aggregation for Federated Learning: Waveform Superposition and Prototype Validation. *J. of Commun. and Inf. Netw.*, 6(4): 429–442.

Jiang, Y.; Konečnỳ, J.; Rush, K.; and Kannan, S. 2019. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*.

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

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2019. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. Apr. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proc. Int. Conf. Artif. Intell. Stat.*, volume 54, 1273–1282. Fort Lauderdale, USA.

Nocedal, J.; and Wright, S. J. 1999. *Numerical optimization*. Springer.

Sery, T.; and Cohen, K. Apr. 2020. On analog gradient descent learning over multiple access fading channels. *IEEE Trans. Signal Process.*, 68: 2897–2911.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. *Advances in neural information processing systems*, 30.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.

Wang, C.; Chen, Z.; Pappas, N.; Yang, H. H.; Quek, T. Q.; and Poor, H. V. 2024. Adaptive Federated Learning Over the Air. *arXiv preprint arXiv:2403.06528*.

Wen, H.; Xing, H.; and Simeone, O. 2024. Pre-Training and Personalized Fine-Tuning via Over-the-Air Federated Meta-Learning: Convergence-Generalization Trade-Offs. *arXiv preprint arXiv:2406.11569*.

Xu, J.; Chen, Z.; Quek, T. Q.; and Chong, K. F. E. 2022. Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10184–10193.

Yang, H. H.; Chen, Z.; Quek, T. Q.; and Poor, H. V. Dec. 2021. Revisiting analog over-the-air machine learning: The blessing and curse of interference. *IEEE J. Sel. Topics Signal Process.*, 16(3): 406–419.

Yang, K.; Jiang, T.; Shi, Y.; and Ding, Z. Mar. 2020. Federated learning via over-the-air computation. *IEEE Trans. Wireless Commun.*, 19(3): 2022–2035.

Zhang, J.; Hua, Y.; Wang, H.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11237–11244.

Appendix

We can write the update of the global parameter in a typical communication round $t+1$ as follows:

$$
\begin{aligned}
&\mathbb{E}\left[F\left(\boldsymbol{w}^{t+1}\right)\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[F\left(\boldsymbol{w}^{t}\right)\right] - \beta\mathbb{E}\left[\left\langle \nabla F\left(\boldsymbol{w}^{t}\right), \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)\right\rangle\right] \\
&+ \frac{\beta^2 L}{2}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right] \\
&= \mathbb{E}\left[F\left(\boldsymbol{w}^{t}\right)\right] + \frac{\beta^2 L}{2}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right] \\
&- \beta\mathbb{E}\left[\left\langle \nabla F\left(\boldsymbol{w}^{t}\right), \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) - \nabla F\left(\boldsymbol{w}^{t}\right) + \nabla F\left(\boldsymbol{w}^{t}\right)\right\rangle\right] \\
&= \mathbb{E}\left[F\left(\boldsymbol{w}^{t}\right)\right] + \frac{\beta^2 L}{2}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right] \\
&+ \beta\mathbb{E}\left[\left\langle -\nabla F\left(\boldsymbol{w}^{t}\right), \frac{1}{N}\sum_{n=1}^{N} \mu_h\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) - \nabla F\left(\boldsymbol{w}^{t}\right)\right\rangle\right] \\
&- \beta\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^{t}\right)\right\|_2^2\right] \\
&\leq \mathbb{E}\left[F\left(\boldsymbol{w}^{t}\right)\right] - \frac{\beta}{2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^{t}\right)\right\|_2^2\right] \\
&+ \underbrace{\frac{\beta}{2}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} \mu_h\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) - \nabla F\left(\boldsymbol{w}^{t}\right)\right\|_2^2\right]}_{Q_1} \\
&+ \underbrace{\frac{\beta^2 L}{2}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right]}_{Q_2}
\end{aligned}
\tag{21}
$$

where (a) follows from Assumption 1 and the fact that $\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}$.

Leveraging the fact that each entry of $\boldsymbol{\xi}_t$ is independent and has a zero mean, we can expand Q_2 as

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\| \frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)\right\|_2^2\right] + \mathbb{E}\left[\left\|\boldsymbol{\xi}_t\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{N}\left(\mu_h^2 + \sigma_h^2\right)\left\|\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)\right\|_2^2\right] + d\sigma_g^2.
\end{aligned}
\tag{22}
$$

Subsequently, we assume $\hat{\boldsymbol{w}}^t$ is a point between $\tilde{\boldsymbol{w}}_n^{t+1}$ and

$\boldsymbol{w}^t$ and apply the Mean Value Theorem to obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right)\right\|_2^2\right] = \mathbb{E}\left[\left\|\nabla f_n(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right))\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\nabla f_n(\boldsymbol{w}^t) - \alpha\nabla^2 f_n\left(\hat{\boldsymbol{w}}^t\right)\nabla f_n(\boldsymbol{w}^t)\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\left(I - \alpha\nabla^2 f_n\left(\hat{\boldsymbol{w}}^t\right)\right)\nabla f_n(\boldsymbol{w}^t)\right\|_2^2\right] \\
&\overset{(a)}{\leq} \left(2 + 2L^2\alpha^2\right)\mathbb{E}[] \\
&= \left(2 + 2L^2\alpha^2\right)\mathbb{E}\left[\left\|\nabla f_n(\boldsymbol{w}^t) + \nabla F\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&\overset{(b)}{\leq} 4\left(1 + L^2\alpha^2\right)\left(G^2 + \sigma^2\right) \qquad (23)
\end{aligned}$$

where (a) follows from Assumption 1, and (b) follows from Assumption 2 and Assumption 3.

By substituting (23) into (22), Q_1 can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N} h_{n,t}\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) + \boldsymbol{\xi}_t\right\|_2^2\right] \\
&\leq \frac{4}{N}\left(\mu_h^2 + \sigma_h^2\right)\left(1 + L^2\alpha^2\right)\left(G^2 + \sigma^2\right) + d\sigma_g^2 \qquad (24)
\end{aligned}$$

For convenience, we define the above expression as C_2.

Next, we bound Q_1 by the following:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\mu_h\nabla f_n\left(\tilde{\boldsymbol{w}}_n^{t+1}\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\|\frac{1}{N}\sum_{n=1}^{N}\mu_h\nabla f_n\left(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right)\right) \\
&- \left(I - \alpha\nabla^2 f_n\left(\boldsymbol{w}^t\right)\right)\nabla f_n\left(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right)\right)\Bigg\|_2^2\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\|\frac{1}{N}\sum_{n=1}^{N}\left(\mu_h - 1\right)\nabla f_n\left(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right)\right) \\
&+ \alpha\nabla^2 f_n\left(\boldsymbol{w}^t\right)\nabla f_n\left(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right)\right)\Bigg\|_2^2\Bigg] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\left(\mu_h - 1\right)\nabla f_n(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right))\right\|_2^2\right] \\
&+ 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\alpha\nabla^2 f_n\left(\boldsymbol{w}^t\right)\nabla f_n\left(\boldsymbol{w}^t - \alpha\nabla f_n\left(\boldsymbol{w}^t\right)\right)\right\|_2^2\right] \\
&\overset{(a)}{\leq} \frac{8}{N}\left(\left(\mu_h - 1\right)^2 + \alpha^2L^2\right)\left(1 + \alpha^2L^2\right)\left(G^2 + \sigma^2\right) \qquad (25)
\end{aligned}$$

where (a) follows from Assumption 1. For convenience, we define the above expression as C_1. To this end, by substituting (24) and (25) into (21), we have:

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{w}^{t+1})\right] \leq\ &\mathbb{E}\left[F\left(\boldsymbol{w}^t\right)\right] - \frac{\beta}{2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&+ \frac{\beta}{2}C_1 + \frac{\beta^2 L}{2}C_2 \qquad (26)
\end{aligned}$$

Finally, by induction we reach:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&\leq \frac{2}{\beta T}\left(\mathbb{E}\left[\nabla F(\boldsymbol{w}^0)\right] - F^*\right) + C_1 + \beta L C_2. \qquad (27)
\end{aligned}$$

The proof is completed.