# Pre-Training and Personalized Fine-Tuning via Over-the-Air Federated Meta-Learning: Convergence-Generalization Trade-Offs

Haifeng Wen, Hong Xing, and Osvaldo Simeone

arXiv:2406.11569v5 [cs.LG] 11 Oct 2025

**Abstract**

For modern artificial intelligence (AI) applications such as large language models (LLMs), the training paradigm has recently shifted to pre-training followed by fine-tuning. Furthermore, owing to dwindling open repositories of data and thanks to efforts to democratize access to AI models, pre-training is expected to increasingly migrate from the current centralized deployments to federated learning (FL) implementations. Meta-learning provides a general framework in which pre-training and fine-tuning can be formalized. Meta-learning-based personalized FL (meta-pFL) moves beyond basic personalization by targeting generalization to new agents and tasks. This paper studies the generalization performance of meta-pFL for a wireless setting in which the agents participating in the pre-training phase, i.e., meta-learning, are connected via a shared wireless channel to the server. Adopting over-the-air computing, we study the trade-off between generalization to new agents and tasks, on the one hand, and convergence, on the other hand. The trade-off arises from the fact that channel impairments may enhance generalization, while degrading convergence. Extensive numerical results validate the theory.

## I. Introduction

For modern artificial intelligence (AI) applications such as large language models (LLMs), the training paradigm has shifted to *pre-training* followed by *fine-tuning* [1]. Pre-training is currently done centrally using large data repositories obtained, often at a high cost, by large corporations [1]–[3]; while fine-tuning is typically much cheaper, and it can be used to *personalize* models

H. Wen and H. Xing are with the IoT Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China; H. Xing is also affiliated with the Department of ECE, The Hong Kong University of Science and Technology, HK SAR (e-mails: hwen904@connect.hkust-gz.edu.cn, hongxing@ust.hk). O. Simeone is with the Department of Engineering, King's College London, London, WC2R 2LS, U.K. (e-mail: osvaldo.simeone@kcl.ac.uk).

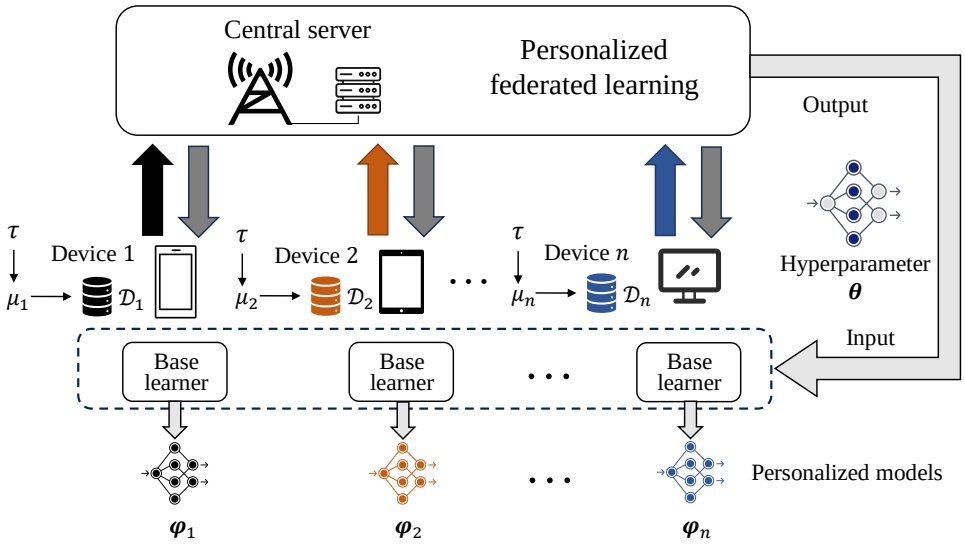



Fig. 1. In the considered personalized federated learning (pFL) setting, each device $i$ aims to find a fine-tuned model $\boldsymbol{\varphi}_i \in \mathbb{R}^d$ using a local data set $\mathcal{D}_i \sim \mu_i^m$ by communicating with the central server. The central server maintains a hyperparameter $\boldsymbol{\theta} \in \mathbb{R}^d$, representing a pre-trained model, which is updated based on information sent by the devices.

to individual requirements [4], [5]. For example, a foundation LLM can be fine-tuned to serve as a personal assistant or as a surrogate CEO based on data tailored to the individual use cases [5]. As large dataset repositories are becoming a scarce commodity, it has become imperative to use distributed data in a privacy-minded way for pre-training and, possibly, fine-tuning [6], [7]. Furthermore, efforts to democratize access to AI models are moving in the direction of federated pre-training implementations, which leverage decentralized computing resources [8].

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 [9], [10]. When applied to a multi-agent, and federated setting, as illustrated in Fig. 1, agents collaborate via communication to a central server for pre-training of a shared model defined by hyperparameters $\boldsymbol{\theta}$, from which each agent $i$ can extract a fine-tuned model $\boldsymbol{\varphi}_i$ that is tailored to the individual data sets at each agent $i$. As in federated learning (FL), agents do not directly exchange data but only model information, thus limiting the leakage of private information [8].

Conventional FL, which aims at finding a single shared model for all agents, is known to suffer from the heterogeneity of the distributions across agents. Heterogeneity has been addressed via methods such as decomposition [11], masking [12], and clustering [13]. *Personalized FL* (pFL)

alleviates the problem of data heterogeneity by optimizing personalized AI models that are adjusted to the local tasks of interest for each user [14].

*Meta-learning-based pFL* (meta-pFL) moves beyond basic personalization by targeting *generalization to new agents and tasks* [15]–[17]. In line with the goal of the pre-training/fine-tuning workflow, meta-pFL does not merely aim for enhancing the performance of existing agents and tasks. Rather, it optimizes shared hyperparameters $\boldsymbol{\theta}$ that can be efficiently personalized via fine-tuning to *new* agents and tasks that are not observed during federated pre-training.

This paper studies the generalization performance of meta-pFL for a wireless setting in which the agents participating in the pre-training phase, i.e., meta-learning, are connected via a shared wireless channel to the server [18], [19]. Wireless meta-pFL systems are suitable for large-scale intelligent Internet of Things (IoT) systems, in which IoT devices collect data for collaborative model training. However, the performance of such systems may suffer from limited access to wireless resources. A popular class of methods has explored the use of over-the-air computing [20], an approach known as *over-the-air FL* (AirFL) [18]–[20], to overcome communication bottlenecks on the shared radio channel for conventional FL.

Channel impairments may not necessarily be deleterious for generalization. In fact, the addition of noise during pre-training can have a regularizing effect [21], and it may even implement approximate forms of Bayesian learning [10] [22]. Accordingly, references such as [23]–[25] have shown improved generalization performance as a result of transmissions of model information over noisy channels in AirFL. The downside of a noisy training process is the convergence time, which may be severely affected by noisy updates [26]–[28]. This paper studies the trade-off between generalization to new agents and tasks, on the one hand, and convergence, on the other hand, for an over-the-air implementation of meta-pFL.

### A. Related Works

AirFL protocols for conventional FL have been widely studied. Among some notable algorithmic contributions, the authors in [19], [29] proposed and optimized truncated-channel-inversion power control strategies, while reference [30] proposed to enhance spectral efficiency via sparsification and linear compression mechanisms. Convergence was studied in [26], [27] for phase compensation-based or channel inversion-based AirFL schemes targeting convex loss functions, showing a linear convergence rate as in the centralized gradient descent (GD). These studies reveal the impact of channel fading, channel noise, and data heterogeneity on the convergence

performance of AirFL. Other related works on FL include [28], [31], which derived and optimized transmission powers to minimize a convergence error upper bound of AirFL.

Several pFL protocols have been proposed that apply in the presence of ideal channels. For instance, some methods leverage knowledge distillation (KD) to obtain personalized models by transferring knowledge of a powerful model to local models [32]; while others divide the trainable model parameters into personalized and shared parts [11]. The meta-pFL protocol studied in [15] builds on the model agnostic meta-learning paradigm (MAML) [33], [34], which applies a two-level gradient descent approach to learn how to quickly adapt to new tasks. Reference [15] analyzed the convergence of meta-pFL in the presence of ideal communication.

Limited work has been done on the design of pFL protocols for wireless systems. Solutions include [35], [36], cluster-based pFL [37], [38] and partial-model-based pFL [39], which are based on digital transmission. The analysis in these papers provides insights into convergence. AirFL-based schemes for pFL were considered in [40]–[42], focusing on convergence analysis and optimal resource allocation.

For meta-pFL, the *generalization error* is a key performance metric, as it quantifies the capacity of a shared model to be efficiently fine-tuned to new tasks. Generalization analyses for conventional AirFL can be found in [25], which shows that heavy-tail noise deteriorates the convergence rate while improving the generalization capacity. References [43], [44] have studied the generalization error for MAML under the assumption of ideal channels. To the best of our knowledge, no analysis of generalization exists for meta-pFL under the assumption of ideal channels, let alone noisy channels.

### *B. Contributions*

This paper investigates a wireless implementation of the meta-pFL scheme that leverages over-the-air computing, with the main goal of understanding the trade-off between convergence and generalization entailed by the wireless transmission of model information. The main contributions are as follows:

1) We introduce a MAML-based over-the-air meta-pFL protocol, which adapts the approach in [15] for use of shared wireless channels. The proposed protocol, termed *Air-meta-pFL*, leverages sparsification and linear compression [30] along with a long-term memory mechanism to compensate for the error caused by the gradient sparsification [45].

2) Convergence bounds are derived for Air-meta-pFL under general smooth and non-convex loss functions with constant and adaptive learning rates. Our results quantify the impact of data heterogeneity, number of active devices, transmit power, number of channel uses, channel fading, and channel noise.
3) An upper bound on the generalization error of Air-meta-pFL is obtained that depends on the mutual information between the model parameters and the training data sets. The derived bound captures the impact of the same factors affecting convergence, namely data heterogeneity, number of active devices, transmit power, number of channel uses, channel fading, and channel noise. As anticipated, the analysis reveals a trade-off between convergence and generalization, with factors enhancing generalization, such as channel noise, potentially impairing convergence.
4) Experimental results corroborate the insights gained from the convergence and generalization bounds, demonstrating the convergence-generalization trade-offs in practical conditions.

The paper is organized as follows. The system model in Section II introduces the original pFL protocol and the communication model. Section III presents the proposed Air-meta-pFL protocol. Section IV and V describe the derived upper bounds on convergence and generalization error, respectively. Numerical results are obtained in VI. Finally, we conclude the paper in VII.

## II. Preliminaries and System Model

In this section, we first review pFL and meta-pFL [15], which operate under the assumption of ideal and noiseless communications. Then, we describe the considered wireless channel model that accounts for fading and noise. Throughout this work, we study the setting in Fig. 1, in which a set $[n] \triangleq \{1, \ldots, n\}$ of devices communicate with an edge server over a multiple access (MAC) fading channel.

### A. Preliminaries

**Personalized Federated Learning (pFL):** As illustrated in Fig. 1, each device $i \in [n]$ is assumed to possess a distinct local data set $\mathcal{D}_i = \{Z_{i,j}\}_{j\in[m]}$, where data point $Z_{i,j}$ with $j \in [m]$ is independent and identically distributed ($i.i.d.$) over a sample space $\mathcal{Z}$. The unknown data-generation distribution for device $i$ is denoted as $\mu_i$, so that each $j$-th data point in data set $\mathcal{D}_i$ is distributed as $Z_{i,j} \sim \mu_i$, and the overall local data set $\mathcal{D}_i$ is distributed as $\mathcal{D}_i \sim \mu_i^m$. The

data distributions $\{\mu_i\}_{i\in[n]}$ of all $n$ devices are drawn $i.i.d.$ from a common distribution $\tau$, i.e., $\mu_i \sim \tau$ for $i \in [n]$. Distribution $\tau$ captures the similarity of the data distribution across devices.

In pFL, the goal of each device $i$ is to minimize the *local test loss*

$$f_i(\boldsymbol{\varphi}_i) = \mathbb{E}_{Z\sim\mu_i}[\ell(\boldsymbol{\varphi}_i; Z)] \tag{1}$$

over a parameter vector $\boldsymbol{\varphi}_i \in \mathbb{R}^d$, where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is a loss function. To this end, using a common vector $\boldsymbol{\theta}$, representing the *pre-trained* model, each device $i \in [n]$ applies a *base learning algorithm* $\mathrm{P}_{\boldsymbol{\varphi}_i|\boldsymbol{\theta},\mathcal{D}_i}$, which implements a, generally stochastic, mapping between the local data set $\mathcal{D}_i$ and the pre-trained vector $\boldsymbol{\theta}$ to a fine-tuned model parameter $\boldsymbol{\varphi}_i$. We denote by

$$F_i(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}_i\sim\mu_i^m}\left[\mathbb{E}_{\boldsymbol{\varphi}_i\sim\mathrm{P}_{\boldsymbol{\varphi}_i|\boldsymbol{\theta},\mathcal{D}_i}}\left[f_i(\boldsymbol{\varphi}_i)\right]\right] \tag{2}$$

the *local meta-test loss* as a function of the pre-trained vector $\boldsymbol{\theta}$. Note that the outer expectation is with respect to (w.r.t.) the distribution $\mu_i^m$ of the local data set $\mathcal{D}_i$, and the inner expectation is over the output of the base learning algorithm.

The pre-trained vector $\boldsymbol{\theta}$ is ideally selected so as to address the problem of minimizing the *meta-test loss* $F(\boldsymbol{\theta})$, which is the average across all devices of the local meta-test loss (2). The corresponding optimization problem is given by

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\texttt{Minimize}} \quad \left\{F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{\theta})\right\}. \tag{3}$$

In order to address problem (3), each device $i \in [n]$ divides the local data set $\mathcal{D}_i$ into two disjoint subsets $\mathcal{D}_i^{(\mathrm{tr})}$ and $\mathcal{D}_i^{(\mathrm{va})}$ such that $\mathcal{D}_i = \mathcal{D}_i^{(\mathrm{tr})} \bigcup \mathcal{D}_i^{(\mathrm{va})}$ with $|\mathcal{D}_i^{(\mathrm{tr})}| = m^{(\mathrm{tr})}$ and $|\mathcal{D}_i^{(\mathrm{va})}| = m^{(\mathrm{va})}$. The training data set $\mathcal{D}_i^{(\mathrm{tr})}$ is utilized to implement the *base learner* $\mathrm{P}_{\boldsymbol{\varphi}_i|\boldsymbol{\theta},\mathcal{D}_i^{(\mathrm{tr})}}$, while the validation data set $\mathcal{D}_i^{(\mathrm{va})}$ is used to estimate the local test loss (1). In particular, the base learning algorithm fine-tunes via one or more steps of (stochastic) gradient descent (GD) the *shared pre-trained vector* $\boldsymbol{\theta} \in \mathbb{R}^d$ using the training data set $\mathcal{D}_i^{(\mathrm{tr})}$. Considering, for simplicity of illustration, a single GD step, the base learner outputs the fine-tuned model

$$\boldsymbol{\varphi}_i = \boldsymbol{\theta} - \frac{\alpha}{m^{(\mathrm{tr})}} \sum_{Z\in\mathcal{D}_i^{(\mathrm{tr})}} \nabla\ell(\boldsymbol{\theta}; Z), \tag{4}$$

where $\alpha > 0$ is the step size. Using the validation data set $\mathcal{D}_i^{(\mathrm{va})}$ to estimate the local test loss (1) yields the following empirical version of problem (3),

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\texttt{Minimize}} \left\{\hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^{(\mathrm{va})}}\sum_{Z'\in\mathcal{D}_i^{(\mathrm{va})}} \ell\left(\boldsymbol{\theta} - \frac{\alpha}{m^{(\mathrm{tr})}}\sum_{Z\in\mathcal{D}_i^{(\mathrm{tr})}}\nabla\ell(\boldsymbol{\theta}; Z); Z'\right)\right\}, \tag{5}$$

where $\mathcal{D}_{1:n} = \bigcup_{i=1}^{n} \mathcal{D}_i$. The objective $\hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta})$ in problem (5) is known as the *meta-training loss*.

**Meta-pFL:** Meta-pFL addresses problem (5) using gradient descent on the shared hyperparameter $\boldsymbol{\theta}$. Specifically, at each communication round $t \in \{0, ..., T-1\}$, a fraction $rn$, with $r \in (0, 1]$, of devices are chosen uniformly at random to transmit to the server. We denote as $\mathcal{I}^{(t)} \subseteq [n]$ the subset of devices *active* at round $t$, which has cardinality $|\mathcal{I}^{(t)}| = rn$. After receiving the current vector $\boldsymbol{\theta}^{(t)}$ at round $t$ from the server, each active device $i \in \mathcal{I}^{(t)}$ initializes the hyperparameter vector $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_i^{(t,0)} = \boldsymbol{\theta}^{(t)}$. Then, it performs $Q$ local SGD steps as

$$\boldsymbol{\theta}_i^{(t,q+1)} = \boldsymbol{\theta}_i^{(t,q)} - \eta^{(t)} \hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}), \tag{6}$$

where $q \in \{0, \ldots, Q-1\}$ is the local-SGD index; $\eta^{(t)}$ denotes the step size at round $t$; and $\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})$ is an estimate of the true gradient $\nabla F_i(\boldsymbol{\theta}_i^{(t,q)})$ of the local meta-test loss (2). This estimate is obtained using mini-batches of device $i$'s local data set $\mathcal{D}_i$ as

$$\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}) = \left(\boldsymbol{I}_d - \alpha \hat{\nabla}^2 f_i(\boldsymbol{\theta}_i^{(t,q)})\right) \hat{\nabla} f_i\left(\boldsymbol{\theta}_i^{(t,q)} - \alpha \hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)})\right), \tag{7}$$

where $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)}) = \frac{1}{|\mathcal{B}_i^{(t)}|} \sum_{Z \in \mathcal{B}_i^{(t)}} \nabla \ell(\boldsymbol{\theta}_i^{(t,q)}; Z)$ is an estimate of the gradient of the local test loss (1) obtained from a mini-batch $\mathcal{B}_i^{(t)} \subseteq \mathcal{D}_i^{(\text{tr})}$ of data points; and the vector $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)} - \alpha \hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)}))$ and Hessian matrix $\hat{\nabla}^2 f_i(\boldsymbol{\theta}_i^{(t,q)})$ are similarly obtained from distinct mini-batches of $\mathcal{D}_i^{(\text{va})}$. Note that the estimate $\hat{\nabla} F_i(\boldsymbol{\theta})$ is not unbiased, i.e., $\mathbb{E}[\hat{\nabla} F_i(\boldsymbol{\theta})] \neq \nabla F_i(\boldsymbol{\theta})$ [15].

Each device $i \in \mathcal{I}^{(t)}$ transmits the hyperparameter model difference

$$\boldsymbol{\Delta}_i^{(t)} = \boldsymbol{\theta}_i^{(t,0)} - \boldsymbol{\theta}_i^{(t,Q)} \tag{8}$$

to the server. The server aggregates the model differences to update the global hyperparameter vector $\boldsymbol{\theta}^{(t+1)}$ as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \frac{1}{rn} \sum_{i \in \mathcal{I}^{(t)}} \boldsymbol{\Delta}_i^{(t)}. \tag{9}$$

The updated $\boldsymbol{\theta}^{(t+1)}$ is then broadcast to all $n$ devices at the beginning of round $t+1$. The above steps are iterated until a convergence criterion is met.

### *B. Communication Model*

The meta-pFL protocol assumes ideal communication between devices and the server. In this subsection, we present a standard over-the-air computing (AirComp) communication model that accounts for limitations imposed by wireless multiple access channels.

At the $t$-th communication round, it takes place over a block of $M$ channel uses. Each active device $i \in \mathcal{I}^{(t)}$ transmits an $M$-dimensional vector $\boldsymbol{x}_i^{(t)}$. The average transmit power $\mathbb{E}||\boldsymbol{x}_i^{(t)}||^2$ must satisfy the constraint as

$$\frac{1}{M}\mathbb{E}\|\boldsymbol{x}_i^{(t)}\|^2 \leq P_i, \tag{10}$$

where the expectation is over the transmitted signal $\boldsymbol{x}_i^{(t)}$, and $\|\cdot\|$ is denoted as $l_2$-norm in vector space. We assume that complex channel $h_i^{(t)} = |h_i^{(t)}|e^{j\phi_i^{(t)}}$ between each device $i \in [n]$ and the server remains constant for the duration of a block, but may vary from block to block. All active devices transmit simultaneously over the MAC, so that the edge server receives an $M \times 1$ vector given by

$$\boldsymbol{y}^{(t)} = \sum_{i \in \mathcal{I}^{(t)}} h_i^{(t)} \boldsymbol{x}_i^{(t)} + \boldsymbol{n}^{(t)}, \tag{11}$$

where $\boldsymbol{n}^{(t)}$ is the additive white Gaussian noise (AWGN) vector with zero mean and variance $\sigma_n^2$. We also assume that the interference from non-AirFL IoT devices is negligible or can be eliminated via signal processing methods [46].

As in [30], we assume ideal and noiseless communication in the downlink, given the less constrained resources available for downlink communication at the edge server.

## III. Over-the-Air Meta-Learning Based Personalized Federated Learning

In this section, we describe the proposed implementation protocol of over-the-air meta-pFL as illustrated in Fig. 2. The analysis of convergence and generalization of the approach will be provided in the next two sections.

### A. *Air-meta-pFL*

As described in the previous section, in meta-pFL at each communication round $t$, each active device $i \in \mathcal{I}^{(t)}$ must communicate its local update $\boldsymbol{\Delta}_i^{(t)}$ in (8) to the server. To this end, as in [26], [28], [30], [47], we adopt an analog communication strategy based on sparsification, linear compression, channel phase compensation, and power scaling.

First, a $k$-contraction operator is applied to sparsify vector $\boldsymbol{\Delta}_i^{(t)}$. The $k$-contraction operator $\text{Comp}_k : \mathbb{R}^d \to \mathbb{R}^d$ preserves only $k$ entries of its input vector, while setting all other entries to zeros, e.g., via a top-$k$ selection or via a random selection strategy [45]. The resulting sparsified update is given by

$$\boldsymbol{g}_i^{(t)} = \text{Comp}_k(\boldsymbol{m}_i^{(t)} + \boldsymbol{\Delta}_i^{(t)}), \tag{12}$$

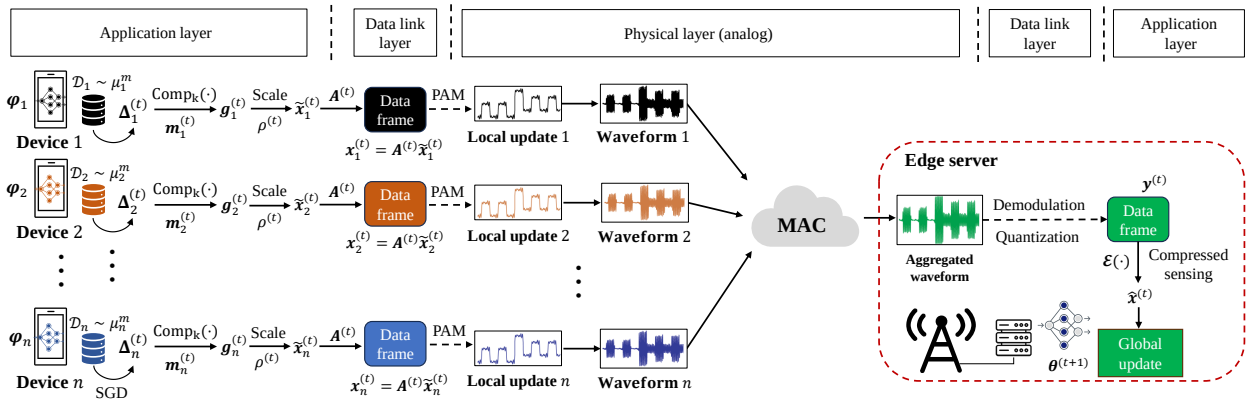



Fig. 2. Illustration of the implementation protocol of over-the-air meta-pFL.

where a memory vector $\boldsymbol{m}_i^{(t)} \in \mathbb{R}^d$ is used to keep track of the accumulated errors as [45]

$$\boldsymbol{m}_i^{(t+1)} = \boldsymbol{m}_i^{(t)} + \left(\boldsymbol{\Delta}_i^{(t)} - \boldsymbol{g}_i^{(t)}\right). \tag{13}$$

The sparsified hyperparameter model difference is scaled as

$$\tilde{\boldsymbol{x}}_i^{(t)} = \frac{\sqrt{\rho^{(t)}} e^{-j\phi_i^{(t)}}}{\eta^{(t)}} \boldsymbol{g}_i^{(t)}, \tag{14}$$

where the factor $e^{-j\phi_i^{(t)}}$ is included to compensate for the phase of channel coefficient $h_i^{(t)} = |h_i^{(t)}| e^{j\phi_i^{(t)}}$, and the common scaling factor $\rho^{(t)}$ is chosen to meet the average power constraint (10).

Finally, in order to reduce the dimensionality of the transmitted vector to fit the $M$ channel uses, each active device $i \in \mathcal{I}^{(t)}$ projects the scaled vector $\tilde{\boldsymbol{x}}_i^{(t)}$ using the same $M \times d$ matrix $\boldsymbol{A}^{(t)}$ as

$$\boldsymbol{x}_i^{(t)} = \boldsymbol{A}^{(t)} \tilde{\boldsymbol{x}}_i^{(t)}. \tag{15}$$

Matrix $\boldsymbol{A}^{(t)}$ is assumed to have the property that its spectral norm $\|\boldsymbol{A}^{(t)}\|_2$, i.e., the square root of the maximum eigenvalue of matrix $\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^T$, satisfies the inequality $\|\boldsymbol{A}^{(t)}\|_2 \leq 1$. This condition guarantees the power constraint (10). Such a compression matrix can be generated by, e.g., selecting $M$ rows of any unitary matrix.

By substituting equation (15) in (11), the vector received at the server can be written as

$$\boldsymbol{y}^{(t)} = \boldsymbol{A}^{(t)} \sum_{i \in \mathcal{I}^{(t)}} h_i^{(t)} \tilde{\boldsymbol{x}}_i^{(t)} + \boldsymbol{n}^{(t)}. \tag{16}$$

The received signal $\boldsymbol{y}^{(t)}$ is used to produce an estimate $\hat{\boldsymbol{x}}^{(t)} = \mathcal{E}(\boldsymbol{y}^{(t)})$ of the weighted sum of updates $\sum_{i\in\mathcal{I}^{(t)}} h_i^{(t)}\tilde{\boldsymbol{x}}_i^{(t)}$, which serves by (12)-(14) as an estimate of the sum of updates $\sum_{i\in\mathcal{I}^{(t)}} \boldsymbol{\Delta}_i^{(t)}$. With such an estimate, the global hyperparameter is finally updated as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \frac{\eta^{(t)}}{\mu_h\sqrt{\rho^{(t)}}rn}\hat{\boldsymbol{x}}^{(t)}, \tag{17}$$

approximating the ideal meta-pFL update (9), where $\mu_h = \mathbb{E}[|h_i^{(t)}|]$ for all $i$ and $t$ (c.f. Assumption 6). Note that the term $\eta^{(t)}/(\mu_h\sqrt{\rho^{(t)}}rn)\hat{\boldsymbol{x}}^{(t)}$ is the unbiased estimate of $1/(rn)\sum_{i\in\mathcal{I}^{(t)}} \boldsymbol{g}_i^{(t)}$. The proposed Air-meta-pFL is summarized in Algorithm 1.

## IV. Convergence Analysis

In this section, we analyze the convergence of the proposed Air-meta-pFL method for general smooth and non-convex loss functions. The main goal is to understand the impact of key parameters such as the signal-to-noise ratio (SNR), the number of devices, data heterogeneity, and the size of the communication block on the number of communication rounds needed for convergence. The next section will highlight the different roles that some of these parameters can play in terms of generalization.

### A. Assumptions

Our analysis is based on the following assumptions.

*Assumption 1 ($k$-Contraction [45]):* For a parameter $0 < k \leq d$, the $k$-contraction operator used in (12) satisfies the inequality

$$\mathbb{E}\|\boldsymbol{x} - \text{Comp}_k(\boldsymbol{x})\|^2 \leq \left(1 - \frac{k}{d}\right)\|\boldsymbol{x}\|^2. \tag{18}$$

*Assumption 2 (Bounded Gradient's Norm, Variance and Hessian's Variance):* For all $i \in [n]$, the gradient and the Hessian of the loss function $\ell(\boldsymbol{\varphi}; Z)$ have bounded variance, i.e.,

$$\mathbb{E}_{Z\sim\mu_i}\|\nabla\ell(\boldsymbol{\varphi}; Z) - \nabla f_i(\boldsymbol{\varphi})\|^2 \leq \sigma_G^2, \quad \mathbb{E}_{Z\sim\mu_i}\left\|\nabla^2\ell(\boldsymbol{\varphi}; Z) - \nabla^2 f_i(\boldsymbol{\varphi})\right\|_2^2 \leq \sigma_H^2, \tag{19}$$

with the norm of the gradient having also a bounded second moment given by

$$\mathbb{E}_{Z\sim\mu_i}\|\nabla\ell(\boldsymbol{\varphi}; Z)\|^2 \leq G^2. \tag{20}$$

*Assumption 3 (Smoothness of the Local Test Loss):* For all $i \in [n]$, the local test loss function $f_i(\cdot)$ is continuously differentiable and $L$-smooth such that for any pair of vectors $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}' \in \mathbb{R}^d$, there exists a constant $L_G > 0$ satisfying the inequality

$$\|\nabla f_i(\boldsymbol{\varphi}) - \nabla f_i(\boldsymbol{\varphi}')\| \leq L_G\|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|. \tag{21}$$

**Algorithm 1:** Air-meta-pFL

**Input:** Learning rate $\{\eta^{(t)}\}_{t=0}^{T-1}$ adaptation learning rate $\alpha$; sizes of minibatches $m_B$ for computing $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)})$, $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)} - \alpha \hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t,q)}))$, and $\hat{\nabla}^2 f_i(\boldsymbol{\theta}_i^{(t,q)})$ in (7); maximum transmit powers $\{P_i\}_{i\in[n]}$; sparsification rate $k/d$; number of local SGD steps $Q$; fraction $r \in (0,1]$ of active devices; and estimator $\mathcal{E}(\cdot)$;

Initialize $t = 0$, $\boldsymbol{\theta}_i^{(0,0)} = \boldsymbol{\theta}^{(0)}$, and $\boldsymbol{m}_i^{(0)} = \mathbf{0}$, for all $i \in [n]$;

**while** $t < T$ **do**

  **Server:**

    Chooses a subset of $rn$ devices $\mathcal{I}^{(t)}$ uniformly at random;

  **end**

  **Active devices** $i \in \mathcal{I}^{(t)}$ **(in parallel):**

    Initialize $\boldsymbol{\theta}_i^{(t,0)} \leftarrow \boldsymbol{\theta}^{(t)}$;

    Perform $Q$ local SGD steps via (6) to obtain the local hyperparameter model difference $\boldsymbol{\Delta}_i^{(t)} = \boldsymbol{\theta}_i^{(t,0)} - \boldsymbol{\theta}_i^{(t,Q)}$;

    Sparsify: $\boldsymbol{g}_i^{(t)} = \mathrm{Comp}_k(\boldsymbol{m}_i^{(t)} + \boldsymbol{\Delta}_i^{(t)})$ ;

    Memory update: $\boldsymbol{m}_i^{(t+1)} = \boldsymbol{m}_i^{(t)} + \boldsymbol{\Delta}_i^{(t)} - \boldsymbol{g}_i^{(t)}$;

    Scale: $\tilde{\boldsymbol{x}}_i^{(t)} = (\sqrt{\rho^{(t)}} e^{-j\phi_i^{(t)}} / \eta^{(t)}) \boldsymbol{g}_i^{(t)}$;

    Transmit $\boldsymbol{x}_i^{(t)} = \boldsymbol{A}^{(t)} \tilde{\boldsymbol{x}}_i^{(t)}$ to the server;

  **end**

  **Server:**

    Receive $\boldsymbol{y}^{(t)} = \boldsymbol{A}^{(t)} \sum_{i\in\mathcal{I}^{(t)}} h_i^{(t)} \tilde{\boldsymbol{x}}_i^{(t)} + \boldsymbol{n}^{(t)}$;

    Estimate: $\hat{\boldsymbol{x}}^{(t)} = \mathcal{E}(\boldsymbol{y}^{(t)})$;

    Global update: $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - (\eta^{(t)} / (\mu_h \sqrt{\rho^{(t)}} rn)) \hat{\boldsymbol{x}}^{(t)}$;

    Broadcast $\boldsymbol{\theta}^{(t+1)}$ to all the $n$ devices;

  **end**

  $t \leftarrow t + 1$;

**end**

**Output:** $\boldsymbol{\theta}^{(T)}$

Furthermore, the Hessian of local test loss function $f_i(\cdot)$ is $L_H$-Lipschitz continuous, i.e., for all pairs $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}' \in \mathbb{R}^d$, we have the inequality

$$\|\nabla^2 f_i(\boldsymbol{\varphi}) - \nabla^2 f_i(\boldsymbol{\varphi}')\|_2 \leq L_H \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\| \tag{22}$$

for some $L_H > 0$.

*Assumption 4 (Bounded Data Heterogeneity [15]):* For any $\boldsymbol{\varphi} \in \mathbb{R}^d$, the gradient $\nabla f_i(\boldsymbol{\varphi})$ of the local test loss and its Hessian $\nabla^2 f_i(\boldsymbol{\varphi})$ satisfy inequalities

$$\|\nabla f_i(\boldsymbol{\varphi}) - \nabla f(\boldsymbol{\varphi})\|^2 \leq \gamma_G^2, \quad \|\nabla^2 f_i(\boldsymbol{\varphi}) - \nabla^2 f(\boldsymbol{\varphi})\|_2^2 \leq \gamma_H^2, \tag{23}$$

where $f(\boldsymbol{\varphi}) = 1/n \sum_{i=1}^n f_i(\boldsymbol{\varphi})$.

*Assumption 5 (Estimation Error):* The estimate $\hat{\boldsymbol{x}}^{(t)} = \mathcal{E}(\boldsymbol{y}^{(t)})$ of the sum of the updates $\sum_{i \in \mathcal{I}^{(t)}} \boldsymbol{\Delta}^{(t)}$ can be expressed as $\hat{\boldsymbol{x}}^{(t)} = \sum_{i \in \mathcal{I}^{(t)}} h_i^{(t)} \tilde{\boldsymbol{x}}_i^{(t)} + \boldsymbol{n}_{\text{est}}^{(t)}$, where the estimation error $\boldsymbol{n}_{\text{est}}^{(t)}$ has zero mean and variance $\mathbb{E}||\boldsymbol{n}_{\text{est}}^{(t)}||^2 = d \cdot v^{(t)}$ for some $v^{(t)} > 0$, and is uncorrelated with the signal $\sum_{i \in \mathcal{I}^{(t)}} h_i^{(t)} \tilde{\boldsymbol{x}}_i^{(t)}$.

Note that Assumption 5 is trivially satisfied for a scheme that does not implement sparsification and linear compression by setting the variance of the estimation error $v^{(t)}$ to be equal to the channel noise power, i.e., $v^{(t)} = \sigma^2$ for all $t$. In the more general case, this assumption is satisfied by applying a linear minimum mean squared error (LMMSE) estimator. It is also approximately met by Bayesian sparse recovery algorithms [48], for which one can evaluate the estimation error $v^{(t)}$ by leveraging state evolution [48], [49].

*Assumption 6 (Time-Varying Channels):* The channel coefficients $h_i^{(t)}$ are $i.i.d.$, with absolute value $|h_i^{(t)}|$ having finite mean $\mathbb{E}[|h_i^{(t)}|] = \mu_h$ and power $\mathbb{E}[|h_i^{(t)}|^2] = \sigma_h^2$.

### *B. Convergence Analysis with Constant Learning Rates*

Under the assumptions listed above, the following convergence result holds for Air-meta-pFL with constant learning rates, i.e., with $\alpha^{(t)} = \alpha$ for the inner SGD update (4) and with $\eta^{(t)} = \eta$ for the outer SGD update (6). As we will see, fixed learning rates yield a convergence rate of the order $\mathcal{O}(1/\sqrt{T})$ in the number of communication rounds $T$, while exhibiting an error floor. The next subsection will show that the error floor can be eliminated via adaptive learning rates, but at the cost of a slower convergence rate.

*Theorem 4.1 (Convergence with Constant Learning Rate):* Under Assumptions 1-6, let $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T-1}$ be the iterates generated by Air-meta-pFL (Algorithm 1) with constant learning rates $\alpha^{(t)} = \alpha \in (0, 1/L_G]$ and $\eta^{(t)} = \eta$ that satisfies the inequalities

$$\eta \in \left\{\eta : \ 0 < \eta \leq \frac{1}{10QL_F}, 60\eta^2 Q^2 L_F^2 + 160\eta^3 Q^3 L_F^3 + 4\eta Q L_F \leq \frac{1}{8}\right\}. \tag{24}$$

Then, on average over the randomness of SGD, sparsification, device selection, fading, and channel noise, Air-meta-pFL satisfies the following inequality

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 \leq \frac{C_0}{\eta T} + \eta(C_n + C_v) + \eta^2(C_F + C_\Lambda) + C_\alpha, \tag{25}$$

where

$$C_0 = \underbrace{\frac{8(F(\boldsymbol{\theta}^{(0)}) - F^*)}{Q}}_{\text{Initialization error}}, \quad C_\alpha = \underbrace{\frac{48\alpha^2 L^2 \sigma_G^2}{m_B}}_{\text{Inner SGD error}}, \quad C_\Lambda = \underbrace{\frac{32Q^2 L_F^2 \Lambda}{r^2}\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right) G^2}_{\text{Sparsification error}},$$

$$C_v = \underbrace{16 L_F Q G^2 (\Lambda + 1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\left(\frac{d}{r^2 n^2 M P_{\min}}\frac{1}{T}\sum_{t=0}^{T-1} v^{(t)} + \frac{2\sigma_h^2}{\mu_h^2} - 2\right)}_{\text{Estimation error}}$$

$$C_n = \underbrace{32QL_F\left(\sigma_F^2 + \gamma_F^2\right)}_{\text{Outer-SGD error \& data heterogeneity}}, \quad C_F = \underbrace{(480Q^2L_F^2 + 1280Q^3L_F^2)(\sigma_F^2 + \gamma_F^2)}_{\text{Outer-SGD error \& data heterogeneity}},$$

where we have defined $\lambda = k/d$, $L_F = 4L + \alpha L_H G$, $\gamma_F^2 = 3G^2\alpha^2\gamma_H^2 + 192\gamma_G^2$, $\Lambda = \frac{(1-\lambda)(1+1/c)}{1-(1-\lambda)(1+c)}$ with $0 < c < \frac{\lambda}{1-\lambda}$, $\sigma_F^2 = 12B^2\sigma_H^2\frac{\alpha^2}{4m_B} + 12\sigma_G^2\left(\frac{1+(\alpha L_G)^2}{m_B}\right)\left(1 + \sigma_H^2\frac{\alpha^2}{4m_B}\right)$, and $P_{\min} = \min_{i\in[n]} P_i$.

*Proof:* The proof is provided in Appendix B. ■

As anticipated, the bound in (25) indicates a convergence to a bounded error with a rate $\mathcal{O}(1/\sqrt{T})$ by setting $\eta$ with the order $\mathcal{O}(1/\sqrt{T})$. The error floor $C_\alpha$ depends on the adaptation learning rate $\alpha$ used in the inner adaptation SGD update (4), as well as on the variance $\sigma_G^2$ of the corresponding gradient. At convergence, the performance is thus limited by the noise in the adaptation step, which can be reduced by increasing the corresponding mini-batch size $m_B$.

The convergence speed, determined by the term scaling as $1/\sqrt{T}$ is dictated by: (*i*) The term $C_0$, accounts for the initialization error; (*ii*) The term $C_n$, which depends on the SGD error on the outer update (6) for the hyperparameter vector, increases with the number $Q$ of outer update steps and with data heterogeneity via parameter $\gamma_F$; and (*iii*) by the term $C_v$, which depends on the variance of the estimation error $v^{(t)}$ due to channel noise, sparsification, and linear compression. This latter term decreases with the number of active devices $rn$, with the

block size $M$, and with the SNR via the power $P_{\min}$, while increasing with the level of data heterogeneity.

The bound (25) also includes a faster contribution that scales as $\mathcal{O}(1/T)$ with $\eta$ set as the order $\mathcal{O}(1/\sqrt{T})$. In a manner similar to the part decreasing as $\mathcal{O}(1/\sqrt{T})$, this contribution is determined by the SGD error on the hyperparameter update, on data heterogeneity, and on the sparsification error. In particular, the term $C_F$ increases with the variance $\sigma_F^2$ of the outer-SGD updates, with the level of data heterogeneity $\gamma_F^2$, and with the number $Q$ of updates; while the term $C_\Lambda$ decreases with $k$, and with the number $Q$ of updates, while decreasing with the mini-batch size $m_B$.

### C. Convergence Analysis with Adaptive Learning Rates

The following Theorem provides an asymptotic analysis of the convergence based on adaptive learning rates. A suitable choice of decreasing learning rates can remove the error floor in (25) due to the noise in the adaptation step, but at the cost of a slower learning rate.

*Theorem 4.2 (Convergence with Adaptive Learning Rate):* Under Assumptions 1-6, let $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T-1}$ be iterates generated according to Air-meta-pFL (Algorithm 1) with an inner learning rate $\alpha^{(t)} = \xi'/(a'+t) \in (0, 1/L_G]$ and an outer learning rate $\eta^{(t)} = \xi/(a+t)$ that satisfies $0 < \eta^{(0)} \leq 1/(10QL_F)$ and $60(\eta^{(0)})^2Q^2L_F^2 + 160(\eta^{(0)})^3Q^3L_F^3 + 4\eta^{(0)}QL_F \leq 1/8$ Then, on average over the randomness of SGD, sparsification, device selection, fading, and channel noise, Air-meta-pFL satisfies the following inequality

$$\min_{t=0,1,\ldots,T-1} \mathbb{E}\|\nabla F(\theta_t)\|^2 \leq \frac{C_{\text{ada}}}{\xi \ln(\frac{T+a-1}{a})}, \tag{26}$$

where

$$\begin{aligned} C_{\text{ada}} = &\underbrace{\frac{8F(\boldsymbol{\theta}^{(0)}) - F^*}{Q}}_{\text{Initialization error}} + \underbrace{\frac{48L_G\sigma_G^2}{m_B}\frac{(\xi')^2}{a'-1}}_{\text{Inner SGD error}} + \underbrace{64Q^2L_F^2G^2\frac{C}{r^2\lambda^2}\left(1+\frac{\sigma_H^2}{L_G^2m_B}\right)\frac{\xi^3}{(a-1)}}_{\text{Sparsification error}} \\ &+ \underbrace{4L_FQG^2\left(1+\frac{\sigma_H^2}{L_G^2m_B}\right)\left(\frac{C}{\lambda^2}+2\right)\frac{\xi^2}{(a-1)}\left(\frac{d}{r^2n^2MP_{\min}}\max_t v^{(t)} + \frac{2\sigma_h^2}{\mu_h^2} - 2\right)}_{\text{Estimation error}} \\ &+ \underbrace{\left(480Q^2L_F^2 + 1280Q^3L_F^2\right)(\sigma_F^2+\gamma_F^2)\frac{\xi^3}{(a-1)^2}}_{\text{Outer-SGD error \& data heterogeneity}} + \underbrace{\frac{32QL_F\xi^2}{(a-1)}\left(\sigma_F^2+\gamma_F^2\right)}_{\text{Outer-SGD error \& data heterogeneity}}, \end{aligned} \tag{27}$$

where $L_F$, $\gamma_F^2$, $\sigma_F^2$, and $P_{\min}$ are defined as in Theorem 4.1, and $C$ is a constant satisfying the inequality $C \geq \frac{4a\lambda(1-\lambda^2)}{a\lambda-4Q}$.

*Proof:* The sketch proof is provided in Appendix B. ■

This result indicates that Air-meta-pFL converges with rate $\mathcal{O}(1/\ln(T))$ when choosing adaptive learning rates as decaying $\mathcal{O}(1/t)$ over the communication round index $t$. The constant $C_{\text{ada}}$ in (26) depends on the same factors as the constants in (25). In particular, the term $C_{\text{ada}}$ decreases with the effective SNR level $P_{\min}/v^{(t)}$, the number of active devices $rn$, and the block size $M$, while increasing with the level of the data heterogeneity $\gamma_F^2$.

## V. Generalization Analysis

Air-meta-pFL aims at finding a shared pre-trained vector $\boldsymbol{\theta}$ that supports the design of effective fine-tuned models $\boldsymbol{\varphi}$ via an SGD-based learner at the device. The previous section took an optimization perspective, addressing the convergence of the protocol to stationary points of the meta-training loss minimization problem (5). However, convergence does not provide any guarantee in terms of test performance. In fact, improving the accuracy of a solution to problem (5) may harm the test performance due to overfitting. To provide a more complete assessment of the performance of Air-meta-pFL, in this section, we analyze the *generalization* performance. As we will argue, communication-related variables, such as received SNR, number of active devices, and number of channel uses, may play opposite roles when viewed through the lenses of convergence and of generalization. This analysis will highlight the emergence of a possible trade-off between convergence and generalization, which will be further studied in Sec. VI via numerical results.

### A. Meta-Generalization Error

As in the generalization analysis of MAML [44], we study the generalization performance of Air-meta-pFL in terms of the local test loss experienced by a new device that did not participate in the federated learning process. Recalling that each device is characterized by a data distribution $\mu \sim \tau$, let us write the corresponding meta-test loss (2) for the new device directly as a function of the data-generation distribution $\mu$ as the expectation

$$F_\mu(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}\sim\mu}\mathbb{E}_{\boldsymbol{\varphi}\sim \mathrm{P}_{\boldsymbol{\varphi}|\boldsymbol{\theta},\mathcal{D}}}\mathbb{E}_{Z\sim\mu}[\ell(\boldsymbol{\varphi}; Z)]. \tag{28}$$

Accordingly, the expected *meta-test loss* for the new device is defined as

$$\bar{F}_\tau(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\mu\sim\tau}[F_\mu(\boldsymbol{\theta})], \tag{29}$$

which is averaged over the possible local distributions $\mu$.

Generalization is measured via the expected gap between meta-training loss $\hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta})$ optimized by meta-pFL as per problem (5) and the target meta-test loss $\bar{F}_{\tau}(\boldsymbol{\theta})$. If this gap is small, then convergence to a solution of problem (5) does indeed enforce a good generalization performance, while this is not the case otherwise. The resulting *meta-generalization error* is defined as

$$\Delta_{\tau} \triangleq \mathbb{E}_{\mathcal{D}_{1:n},\boldsymbol{\theta}}[\bar{F}_{\tau}(\boldsymbol{\theta}) - \hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta})], \tag{30}$$

where the average is over the distribution of the global data set $\mathcal{D}_{1:n}$ and of the pre-trained vector $\boldsymbol{\theta}$ produced by the federated learning process based on the global data set $\mathcal{D}_{1:n}$. We write the conditioned distribution of the pre-trained vector given the global data set $\mathcal{D}_{1:n}$ as $\mathrm{P}_{\boldsymbol{\theta}|\mathcal{D}_{1:n}}$. The conditioned distribution $\mathrm{P}_{\boldsymbol{\theta}|\mathcal{D}_{1:n}}$ describe the overall operation of the protocol.

An upper bound on the meta-generalization error $\Delta_{\tau}$ in (30) would provide a measure of the discrepancy between the meta-training loss in (5), which is optimized by the protocol, and the target meta-test loss. In light of this, in the following subsections, we aim to analyze the generalization of the proposed Air-meta-pFL by deriving an upper bound of the error defined in (30).

### *B. Assumptions*

The generalization analysis requires the following additional assumptions.

*Assumption 7 (Sub-Gaussian Per-task Training Loss [43]):* For all hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^d$, the *per-task training loss* function

$$L(\boldsymbol{\theta}|\mathcal{D}) \triangleq \frac{1}{m^{(\mathrm{va})}} \sum_{Z \in \mathcal{D}^{(\mathrm{va})}} \ell\left(\boldsymbol{\theta} - \frac{\alpha}{m^{(\mathrm{tr})}} \sum_{Z \in \mathcal{D}^{(\mathrm{tr})}} \nabla \ell(\boldsymbol{\theta}; Z); Z\right) \tag{31}$$

is $\sigma^2$-sub-Gaussian, where randomness is due to the dependence on the local data set $\mathcal{D} \sim \mu^m$.

Assumption 7 is automatically satisfied with $\sigma^2 = (b-a)^2/4$ if the loss function $\ell(\cdot;\cdot)$ satisfies the inequality $a \leq \ell(\cdot;\cdot) \leq b$. In practice, one can always satisfy the sub-Gaussian assumption by clipping the original loss or by applying a composition with a squashing function such as sigmoid function [50].

*Assumption 8 (Independent Mini-Batches):* The mini-batch sampling strategy for calculating the SGD in (6) is such that the selected mini-batch is independent of the hyperparameter $\boldsymbol{\theta}_i^{(t,q)}$, and of mini-batches of previous rounds for all $t \in \{0, 1, ..., T-1\}$, $q \in \{0, 1, ..., Q-1\}$ and all devices $i \in [n]$.

*Assumption 9 (Bounded Sparsified Updates):* For all $i \in [n]$ and $t \in \{0, 1, ..., T-1\}$, there exists a constant $\epsilon_g > 0$ such that the expected square norm of the sparsified update (12) satisfies the inequality $\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2 \geq \eta^2 \epsilon_g$, where $\eta$ is the learning rate in (6).

### C. Generalization Analysis

To start, we review the following lemma derived in [43], which bounds the meta-generalization error in (30) for any protocol described by a conditioned distribution $\mathrm{P}_{\boldsymbol{\theta}|\mathcal{D}_{1:n}}$ via the mutual information $I(\boldsymbol{\theta}; \mathcal{D}_{1:n})$ between the pre-trained vector $\boldsymbol{\theta}$ and the training data set $\mathcal{D}_{1:n}$. Note that conventional techniques to analyze generalization performance, such as VC dimension or Rademacher complexity, yield bounds that are determined by the complexity of the model class rather than training algorithms. By comparison, the mutual information is algorithm-dependent [44], yielding stronger design insights.

*Lemma 5.1 (Theorem 1, [43]):* Under Assumption 7, for any protocol, characterized by a conditional distribution $\mathrm{P}_{\boldsymbol{\theta}|\mathcal{D}_{1:n}}$, the meta-generalization error $\Delta_\tau$ in (30) satisfies the inequality

$$|\Delta_\tau| \leq \sqrt{\frac{2\sigma^2}{n} I(\boldsymbol{\theta}; \mathcal{D}_{1:n})}. \tag{32}$$

Lemma 5.1 suggests that a smaller correlation between the pre-trained vector $\boldsymbol{\theta}$ and training data sets $\mathcal{D}_{1:n}$, and thus smaller mutual information $I(\boldsymbol{\theta}; \mathcal{D}_{1:n})$, improve the generalization performance. This is because a smaller correlation entails a more limited sensitivity to the specific realization of the training data.

In the sequel, we derive such a specific form, revealing the unique role played by the communication parameters in generalization that distinguishes itself from in convergence.

*Theorem 5.1 (Generalization of Air-meta-pFL):* Under Assumptions 2, 3, 5, 7, 8 and 9, the meta-generalization error of Air-meta-pFL satisfies the following inequality

$$|\Delta_\tau| \leq \sqrt{\frac{d\sigma^2}{n} \sum_{t=0}^{T-1} \log\left(1 + \frac{M P_{\max} r n C_g \sum_{i \in \mathcal{I}^{(t)}} |h_i^{(t)}|^2}{d v^{(t)} \epsilon_g}\right)}, \tag{33}$$

where $C_g = 4Q^2G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)$, $P_{\max} = \max_{i\in[n]} P_i$, and $\Lambda = \frac{(1-\lambda)(1+1/c)}{1-(1-\lambda)(1+c)}$ with $\lambda = k/d$ and $0 < c < \frac{\lambda}{1-\lambda}$.

*Proof:* See Appendix C. ■

*Remark 5.1:* To bound the mutual information $I(\boldsymbol{\theta}; \mathcal{D}_{1:n})$, we need to bound the conditional entropy $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)})$ by data processing inequality (c.f. (62), Appendix C). Conditioned on

$\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)}$, $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)})$ is bounded by the entropy of a Gaussian random variable with the same covariance, i.e., $\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)}$. As a result, the key step remains to bound the norm square $\mathbb{E}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)}\|^2$. However, this is technically challenging as it involves bounding algorithm-specific parameters including the square norm of the mini-batch gradient $\mathbb{E}\|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\|^2$, the memory vector $\mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2$ and the power scaling factor $\rho^{(t)}$ (c.f. (67), Appendix C), none of which is covered by general results provided in [42] and [43] and thus requires exclusive analysis via Lemma A.5, A.7 and A.8.

The bound in (33), while generally not tight, provides insights into the impact of the number of active devices, of SGD, of the available communication resources, and of data heterogeneity on generalization. On one hand, like the convergence bound (25), the generalization bound in (33) increases with the variance $\sigma_F^2$ of the outer stochastic gradient, as well as with the level of data heterogeneity via the parameter $\gamma_F^2$. However, on the other hand, the impact of the communication parameters on generalization is significantly different from that revealed on convergence.

For example, increasing the number of channel uses $M$ compromises generalization, while improving on convergence performance by reducing estimation error (c.f. (25)). Meanwhile, setting $P_{\min} = P_{\max} = P$ for simplicity, the generalization bound in (33) demonstrates that a small SNR $P|h_i^{(t)}|^2/v^{(t)}$ decreases the generalization bound at the cost of deteriorating the convergence via estimation error. Intuitively, this result stems from the fact that an increased disturbance on the communication channel decreases the correlation between training data and pre-trained vector $\boldsymbol{\theta}$, possibly reducing overfitting.

## VI. Numerical Results

In this section, numerical experiments are conducted to validate the performance of the proposed Air-meta-pFL protocol, as well as the convergence and generalization analysis presented in the previous sections.

### A. Experimental Settings

We adopt the Omniglot data set, which contains $1643$ characters, considered as classes, with each class having $p = 20$ instances drawn by different persons, amounting to $32860$ instances in total [33], [44], [51]. We employ $n = 9$ devices for pFL training and $n_{\text{test}} = 3$ additional devices for evaluation of the generalization performance. All devices are assigned non-$i.i.d.$ local data sets.

TABLE I

EXPERIMENTAL PARAMETERS

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| Mini-batch size | $m_B = 32$ | $N$-way | $N = 5$ |
| Local SGD | $Q = 5$ for convergence analysis<br>$Q = 1$ for generalization analysis | $K$-shot | $K = 8$ |
| Number of devices for training | $n = 9$ | Compression rate | $M/d = 0.4$ |
| Number of devices for test | $n_{\text{test}} = 3$ | Sparsification rate | $k/d = 0.04$ |
| Fraction of active device | $r = 0.3$ for convergence analysis<br>$r = 1$ for generalization analysis | Linear compression matrix $\boldsymbol{A}$ | $M$ out of $d$ rows DFT matrix |
| Number of global rounds | $T = 200$ | Channel fading | Rayleigh fading |
| Meta-learning rate | $\eta^{(t)} = \eta = 0.4$ | Estimator $\mathcal{E}(\cdot)$ | OAMP with 20 iterations [48] |
| Inner learning rate | $\alpha^{(t)} = \alpha = 0.4$ | Max. # Monte Carlo trials | 100 |

The experimental setting is illustrated in Fig. 3. Specifically, our experiments adopt the $N$-way $K$-shot classification protocol as follows [33]. Note that first-order MAML achieves comparable test accuracy as MAML for $N$-way $K$-shot tasks [33, Table 1] (see Fig. 6), which provides evidence that the proposed method with first-order approximation is potentially applicable to large models. The task of each training device $i \in [n]$ is to distinguish among $N$ different classes, selected out of a set of $0 \leq m_c \leq 1643$ classes specific to each device, given $K$ instances of each of the $N$ classes. Performance is then assessed by the test accuracy for classification of instances on newly $N$ selected classes within the device-specific $m_c$ classes. For each training device $i \in [n]$, $m_c$ different characters (classes) are drawn uniformly without replacement from Omniglot. The data set $\mathcal{D}_i$ includes $p = 20$ instances for each of the $m_c$ classes. The local data sets for the $n_{\text{test}}$ test devices are generated following the same procedure. Accordingly, the level of data heterogeneity across devices decreases with $m_c$. In our experiment, we set $N = 5$ and $K = 8$. At each communication round $t$, each active device $i \in \mathcal{I}^{(t)}$ randomly chooses a mini-batch $m_B = 32$ $N$-way $K$-shot tasks from data set $\mathcal{D}_i$ for inner update and another $m_B = 32$ tasks for outer update. The tasks are obtained by selecting $N$ classes uniformly at random from the $m_c$ available classes. The meta-test loss $\bar{F}_\tau(\boldsymbol{\theta})$ and test accuracy are averaged over 1000 different $N$-way $K$-shot tasks of each test device, and over $n_{\text{test}}$ test devices.

We employed a CNN network architecture comprising several modules. The initial three modules share identical structures, each featuring a $3 \times 3$ 2D convolution layer with 64 filters

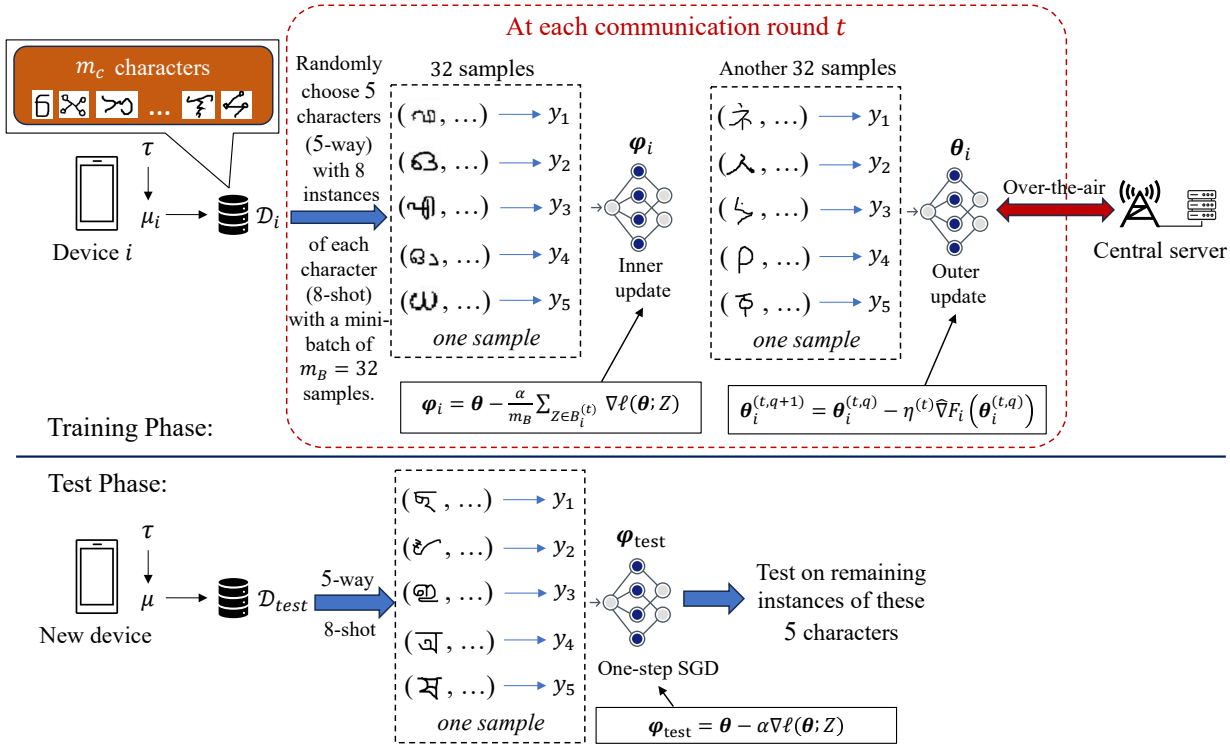



Fig. 3. Illustration of the experimental setting for the Omniglot data set ($N = 5, K = 8, m_B = 32, m_c = 136$ or $10$).

and a stride of 2, followed by a ReLU activation layer and batch normalization (BN). The fourth module consists of a $2 \times 2$ 2D convolution layer with 64 filters and a stride of 1, along with a Relu activation layer and BN. This feature map with shape $64 \times 1 \times 1$ is then fed into a fully connected layer with $5$ nodes. This CNN model results in $d = 91781$ parameters.

For the communication setup, we consider the block-flat Rayleigh fading channel, i.e., $h_i^{(t)} \sim \mathcal{CN}(0, 1)$ in (11), and control the received SNR to be approximately equal to $19$ dB. The number of available channel uses for each device is set as $M = 0.4d$. The linear compression matrix $\boldsymbol{A}^{(t)}$ is generated by randomly selecting $M$ out of $d$ rows of a $d$-by-$d$ discrete Fourier transform (DFT) matrix. Unless otherwise noted, the experimental parameters of the Air-meta-pFL are summarized in Table I.

### *B. Convergence Analysis*

We first study the convergence to stationary solution of the meta-training loss by evaluating the average square norm gradients of the meta-training loss, $1/T\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2$, versus the communication round $T$ with two different values of $m_c$ in Fig. 4. We refer to this quantity as the

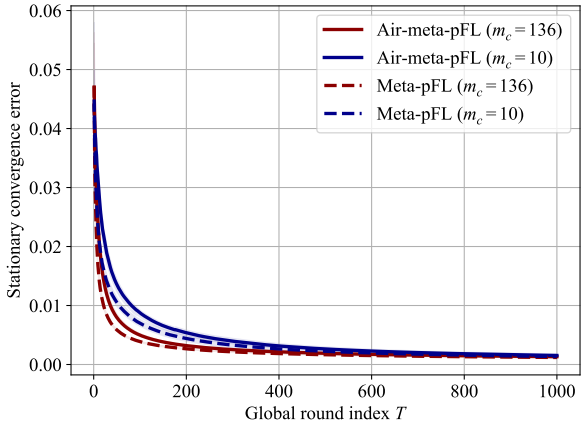



Fig. 4. Stationary convergence error, i.e., square norm of the gradient of the meta-training loss, versus global round $t$ for meta-pFL, which assumes ideal communication, and for Air-meta-pFL. The shaded error bars correspond to intervals covering 95% of the realized values, obtained from the 10 Monte Carlo trials.

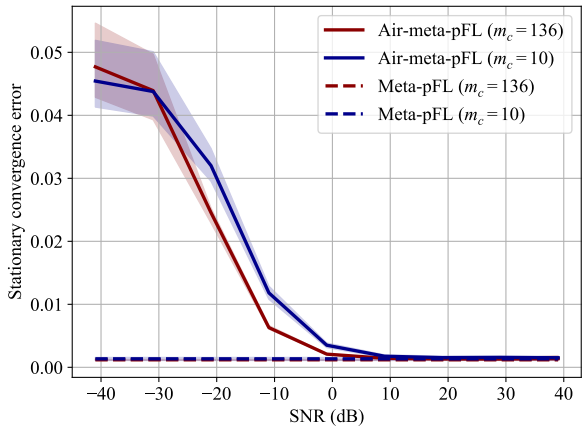



Fig. 5. Stationary convergence error versus received SNR for Air-meta-pFL and meta-pFL.

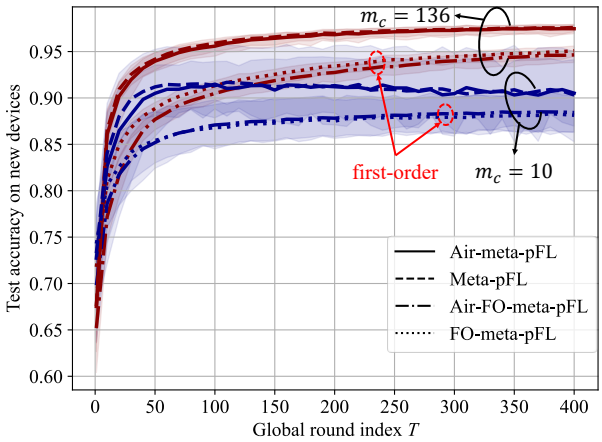



Fig. 6. Test accuracy versus global round $t$ for Air-meta-pFL, meta-pFL and their first-order variants, Air-FO-meta-pFL and FO-meta-pFL.

stationary convergence error. It is observed that the stationary convergence error of Air-meta-pFL decreases to a value similar to PerFedAvg, which assumes ideal communications. Moreover, as predicted by Theorem 4.1, a larger heterogeneity, i.e., smaller $m_c$, leads to a slower convergence.

We further evaluate the stationary convergence error as a function of the received SNR in Fig. 5. In line with Theorem 4.1, the error decreases with the SNR, yielding the same performance as the idealized meta-pFL scheme.

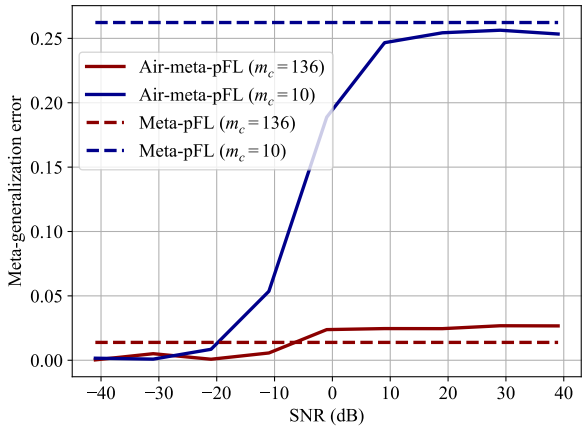



Fig. 7. Meta-generalization error $|\Delta_\tau|$ versus the received SNR for Air-meta-pFL and meta-pFL.

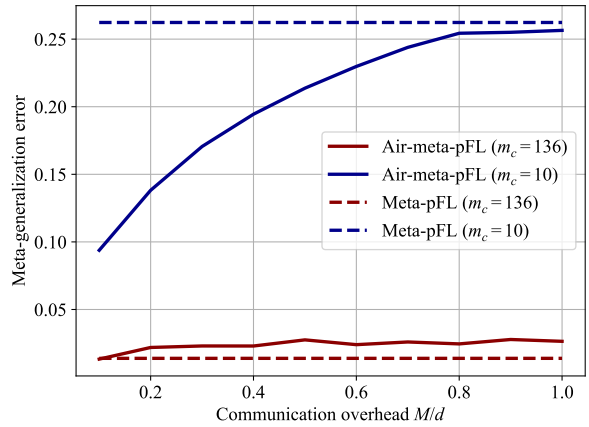


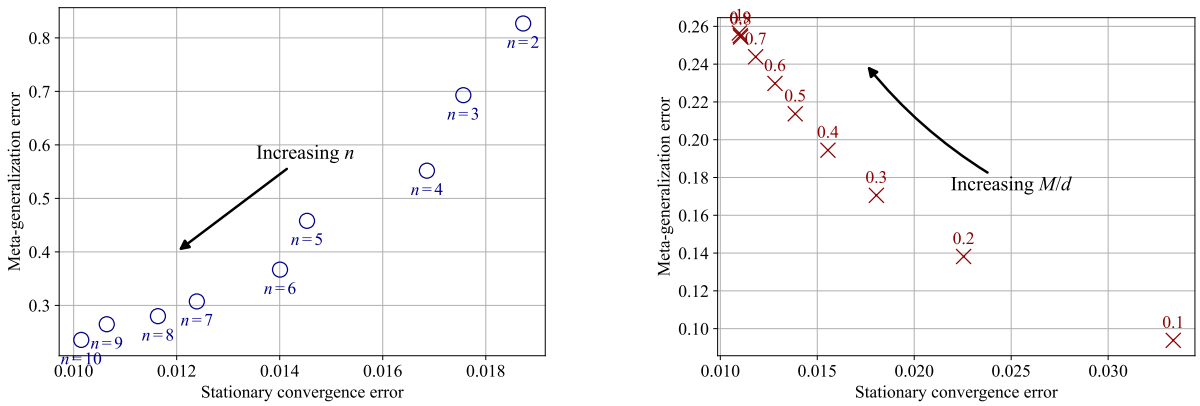

Fig. 8. Meta-generalization error $|\Delta_\tau|$ versus communication overhead $M/d$ with compression rate equal to $k/d = 0.06$ for Air-meta-pFL and meta-pFL.

Finally, we compare the test accuracy on the test devices for Air-meta-pFL, meta-pFL, and their first-order variants, Air-FO-meta-pFL and FO-meta-pFL, versus the communication round $T$ in Fig. 6. Air-meta-pFL is observed to achieve performance comparable to meta-pFL despite the limitations imposed by wireless communication. As suggested by the generalization bound and the convergence bound, Fig. 6 shows a large heterogeneity level, i.e., a smaller $m_c$, which deteriorates the learning performance. Furthermore, Fig. 6 reveals that the test accuracy degradation induced by the first-order approximation is less than $3\%$ compared to the original algorithms.

### C. Generalization Analysis

Next, we conduct experiments to validate the insights obtained by the generalization analysis in Theorem 5.1. To this end, the meta-generalization error $\Delta_\tau$ in (30) is calculated as the difference between the expected meta-test loss $\mathbb{E}_{\boldsymbol{\theta}}[\bar{F}_\tau(\boldsymbol{\theta})]$ and the expected meta-training loss $\mathbb{E}_{\mathcal{D}_{1:n},\boldsymbol{\theta}}[\hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta})]$. The meta-training loss $\hat{F}_{\mathcal{D}_{1:n}}(\boldsymbol{\theta})$ and the meta-test loss $\bar{F}_\tau(\boldsymbol{\theta})$ are both averaged over the shared parameter $\boldsymbol{\theta}$ via Monte Carlo trials.

We focus on the impact of communication impairments on the generalization error $\Delta_\tau$ for different levels of heterogeneity. Fig. 7 shows the absolute value of meta-generalization error $|\Delta_\tau|$ versus the received SNR, with the ratio between the number of preserved parameters via top-$k$ and the number of parameters given by $k/d = 0.06$ and the ratio between the number of




(a) (b)

Fig. 9. Generalization-convergence trade-off by varying: (a) the number of training devices $n$; and (b) the communication overhead $M/d$ with $m_c = 10$ averaged over 10 Monte Carlo trials.

channel uses and the number of parameters given by $M/d = 0.8$. This result shows that, in line with Theorem 5.1, increasing the received SNR entails a larger meta-generalization error $\Delta_\tau$, which also increases with the level of heterogeneity. Thus, communication noise will potentially mitigate overfitting. In fact, for a large $m_c$, Air-meta-pFL may even outperform meta-pFL in terms of generalization, owing to the beneficial effects of channel noise.

Fig. 8 shows the absolute value of meta-generalization error, $|\Delta_\tau|$, versus the communication overhead rate $M/d$ with $k/d = 0.06$. Confirming the prediction of Theorem 5.1, increasing $M/d$ leads to smaller estimation errors, allowing Air-meta-pFL to approach, and potentially surpass, the performance of meta-pFL.

### *D. Convergence-Generalization Trade-Off*

In this subsection, we finally conduct experiments to elucidate the convergence-generalization trade-off. To this end, Fig. 9 shows the meta-generalization error, $|\Delta_\tau|$, versus the stationary convergence error, $1/T\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2$, where we set $m_c = 10$. In Fig. 9(a), the trade-off is traced by varying the number of training devices $n$, while in Fig. 9(b), we vary the communication overhead $M/d$. From Fig. 9(a), as indicated by Theorem 4.1 and Theorem 5.1, increasing the number of training devices $n$ benefits both generalization and convergence. As shown in Table II, the test accuracy is seen to increase with the number of clients $n$, demonstrating the scalability and efficiency of the proposed Air-meta-pFL. In contrast, from Fig. 9(b), as also

reflected by the theorems, the meta-generalization error increases, while the stationary convergence error decreases with $M/d$. This highlights the different roles played by the communication parameters on convergence and generalization. In this regard, communication impairments have a regularizing effect on generalization, while negatively affecting convergence.

TABLE II

TEST ACCURACY VERSUS DIFFERENT NUMBERS OF DEVICES $n$ FOR OMNIGLOT

| $n$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| Test accuracy | 0.855 | 0.882 | 0.887 | 0.901 | 0.906 | 0.919 | 0.919 | 0.922 | 0.923 |

## VII. CONCLUSION AND DISCUSSION

In this paper, we studied a wireless implementation of meta-pFL, a formalized framework of customizing pre-trained models to new agents and tasks using distributed data sets in a federated way. Aiming for efficient use of the shared wireless resources, we introduced a wireless protocol that leverages over-the-air computing, model sparsification, and linear compression, along with an error-compensation mechanism. The main result of this paper is the derivation of bounds on convergence and generalization error, with generalization referring to the performance upon fine-tuning on new users and tasks. The analysis offers insights into the trade-off between convergence and generalization, as wireless impairments may compromise convergence, while potentially enhancing generalization. For example, a larger transmission power leads to convergence speedups, while potentially degrading generalization performance.

Future work will investigate these trade-offs in scenarios involving more complicated training algorithms such as fully decentralized implementations, asynchronous updates, and fine-tuning of LLMs, as well as large action models (LAMs). For instance, recent advancements in orthogonal time frequency and space (OTFS) modulation potentially improve performance in dynamic environments like linear time-varying (LTV) channels [52]. Another interesting future direction is to optimize transmit power and the number of channel uses to improve on the convergence-generalization trade-offs.

## APPENDIX

### A. Important Lemmas

In this subsection we provide several important lemmas for the proofs.

*Lemma A.1 (Young's inequality):* For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ and $\epsilon > 0$, we have

$$|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq \frac{\epsilon}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2\epsilon}\|\boldsymbol{b}\|^2.$$

*Lemma A.2 (Lemma 4.2 in [15]):* Suppose that Assumption 2 and 3 are satisfied and $\alpha \in (0, 1/L_G]$, the local loss $F_i(\boldsymbol{\theta}) \triangleq f_i(\boldsymbol{\theta} - \alpha \nabla f_i(\boldsymbol{\theta}))$ is smooth with parameter $L_F = 4L + \alpha L_H G$.

*Lemma A.3 (Lemma 4.3 in [15]):* The estimate $\hat{\nabla} F_i(\boldsymbol{\theta})$ (c.f. (7)) is computed using independent batches with the same size $m_B$. Suppose that Assumption 2 and 3 satisfied, then for any $\alpha \in (0, 1/L_G]$, $i \in [n]$ and $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\left\| \mathbb{E}\left[ \hat{\nabla} F_i(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w}) \right] \right\|^2 \leq \frac{4\alpha^2 L_G^2 \sigma_G^2}{m_B},$$

and

$$\mathbb{E}\left[ \left\| \hat{\nabla} F_i(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w}) \right\|^2 \right] \leq \sigma_F^2 \triangleq 12\sigma_G^2 \left( \frac{1}{m_B} + \frac{(\alpha L_G)^2}{m_B} \right) \left( 1 + \sigma_H^2 \frac{\alpha^2}{4m_B} \right) + 12 G^2 \sigma_H^2 \frac{\alpha^2}{4m_B},$$

where the expectation is over the randomness of SGD.

*Lemma A.4 (Lemma 4.4 in [15]):* Suppose that Assumptions 2-4 are satisfied, for any $\alpha \in (0, 1/L_G]$ and $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\|\nabla F_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|^2 \leq \gamma_F^2 := 3G^2\alpha^2\gamma_H^2 + 192\gamma_G^2.$$

*Lemma A.5:* For all $i \in [n]$, $t \in \{0, 1, \ldots, T-1\}$ and $q \in \{0, 1, \ldots, Q-1\}$, with independent mini-batches $\mathcal{B}_i^{(t,q)}, \mathcal{B}_i'^{(t,q)}$, and $\mathcal{B}_i''^{(t,q)} \subseteq \mathcal{D}_i$ of the same size $m_B$, the expected square norm of $\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})$ is bounded by

$$\mathbb{E}\|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\|^2 \leq 2\left( (1 + \alpha L_G)^2 + \frac{\alpha^2 \sigma_H^2}{m_B} \right) G^2.$$

*Proof:* For any $i \in [n]$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$\begin{aligned} \|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\| &= \left\| (\boldsymbol{I}_d - \alpha \nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)}) + \boldsymbol{e}_i) \hat{\nabla} f_i\left( \boldsymbol{\theta}_i^{(t,q)} - \alpha \nabla f_i(\boldsymbol{\theta}_i^{(t,q)}, \mathcal{B}_i^{(t,q)}), \mathcal{B}_i'^{(t,q)} \right) \right\| \\ &\leq \|\boldsymbol{I}_d - \alpha \nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)}) + \boldsymbol{e}_i\|_2 \left\| \hat{\nabla} f_i\left( \boldsymbol{\theta}_i^{(t,q)} - \alpha \nabla f_i(\boldsymbol{\theta}_i^{(t,q)}, \mathcal{B}_i^{(t,q)}), \mathcal{B}_i'^{(t,q)} \right) \right\|, \end{aligned} \tag{34}$$

where $\boldsymbol{e}_i = \alpha(\nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)}) - \hat{\nabla}^2 f_i(\boldsymbol{\theta}_i^{(t,q)}, \mathcal{B}_i''^{(t,q)}))$. By Assumption 2, $\mathbb{E}\|\boldsymbol{e}_i\|_2^2 \leq \alpha^2\sigma_H^2/m_B$. Then we have,

$$\mathbb{E}\|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\|^2 \tag{35}$$

$$\begin{aligned} &\overset{(a)}{\leq} \mathbb{E}\left[ \left( 2\|\boldsymbol{I}_d - \alpha \nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)})\|_2^2 + 2\|\boldsymbol{e}_i\|_2^2 \right) \left\| \hat{\nabla} f_i\left( \boldsymbol{\theta}_i^{(t,q)} - \alpha \nabla f_i(\boldsymbol{\theta}_i^{(t,q)}, \mathcal{B}_i^{(t,q)}), \mathcal{B}_i'^{(t,q)} \right) \right\|^2 \right] \\ &\overset{(b)}{=} \left( 2\|\boldsymbol{I}_d - \alpha \nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)})\|^2 + 2\mathbb{E}\|\boldsymbol{e}_i\|^2 \right) \mathbb{E}\left\| \hat{\nabla} f_i\left( \boldsymbol{\theta}_i^{(t,q)} - \alpha \nabla f_i(\boldsymbol{\theta}_i^{(t,q)}, \mathcal{B}_i^{(t,q)}), \mathcal{B}_i'^{(t,q)} \right) \right\|^2 \\ &\overset{(c)}{\leq} 2\left( (1 + \alpha L_G)^2 + \frac{\alpha^2 \sigma_H^2}{m_B} \right) G^2, \end{aligned} \tag{36}$$

where (a) is followed by Jensen's inequality, (b) is due to the independence between $\mathcal{B}_i^{(t,q)}, \mathcal{B}_i^{\prime(t,q)}$, and $\mathcal{B}_i^{\prime\prime(t,q)}$, (c) is follwed by Assumption 2 and 3, and $\|\boldsymbol{I}_d - \alpha\nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)})\|_2 \leq \|\boldsymbol{I}_d\|_2 + \|\alpha\nabla^2 f_i(\boldsymbol{\theta}_i^{(t,q)})\|_2 \leq 1 + \alpha L_G$. ■

*Lemma A.6:* For $\eta \leq \frac{1}{10QL_F}$, we have

$$\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q)} - \boldsymbol{\theta}^{(t)}\|^2 \leq 40Q^2\eta^2(\sigma_F^2 + \gamma_F^2) + 40Q^2\eta^2\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2.$$

*Proof:*

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q)} - \boldsymbol{\theta}^{(t)}\|^2 \qquad (37)\\
&= \mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)} - \eta\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q-1)})\|^2\\
&\overset{(a)}{\leq} \left(1 + \frac{1}{2Q-1}\right)\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)}\|^2 + 2Q\mathbb{E}\|\eta\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q-1)})\|^2\\
&= \left(1 + \frac{1}{2Q-1}\right)\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)}\|^2 + 2Q\eta^2\mathbb{E}\Big\|\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q-1)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q-1)}) + \nabla F_i(\boldsymbol{\theta}_i^{(t,q-1)})\\
&\qquad + \nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F_i(\boldsymbol{\theta}^{(t)}) + \nabla F(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\Big\|^2\\
&\overset{(b)}{\leq} \left(1 + \frac{1}{2Q-1}\right)\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)}\|^2 + 8Q\eta^2\mathbb{E}\|\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q-1)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q-1)})\|^2\\
&\qquad + 8Q\eta^2\Big(\mathbb{E}\|\nabla F_i(\boldsymbol{\theta}_i^{(t,q-1)}) - \nabla F_i(\boldsymbol{\theta}^{(t)})\|^2 + \|\nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\|^2 + \|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\Big)\\
&\overset{(c)}{\leq} \left(1 + \frac{1}{2Q-1} + 8Q\eta^2L_F^2\right)\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)}\|^2 + 8Q\eta^2\sigma_F^2 + 8Q\eta^2\gamma_F^2 + 8Q\eta^2\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2, \qquad (38)
\end{aligned}$$

where (a) and (b) are followed by Lemma A.1, (c) is due to Lemma A.2, Lemma A.3 and Lemma A.4. Let $\eta \leq \frac{1}{10QL_F}$, which implies $\frac{1}{2Q-1} + \frac{8}{100Q} \leq \frac{1}{Q-1}$ for any $Q > 1$. Then, we have

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q)} - \boldsymbol{\theta}^{(t)}\|^2\\
&\leq \left(1 + \frac{1}{Q-1}\right)\mathbb{E}\|\boldsymbol{\theta}_i^{(t,q-1)} - \boldsymbol{\theta}^{(t)}\|^2 + 8Q\eta^2\sigma_F^2 + 8Q\eta^2\gamma_F^2 + 8Q\eta^2\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\\
&\leq \sum_{q=0}^{Q-1}\left(1 + \frac{1}{Q-1}\right)^q\left(8Q\eta^2(\sigma_F^2 + \gamma_F^2) + 8Q\eta^2\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\right)\\
&\leq (Q-1)\left[\left(1 + \frac{1}{Q-1}\right)^Q - 1\right]\Big(8Q\eta^2(\sigma_F^2 + \gamma_F^2) + 8Q\eta^2\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\Big)\\
&\leq 40Q^2\eta^2(\sigma_F^2 + \gamma_F^2) + 40Q^2\eta^2\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2, \qquad (39)
\end{aligned}$$

where the last inequality is due to $(1 + \frac{1}{Q-1})^Q \leq 5$ for all $Q > 1$. When $Q = 1$, the bound is trivially satisfied since $\boldsymbol{\theta}_i^{(t,Q-1)} = \boldsymbol{\theta}_i^{(t,0)} = \boldsymbol{\theta}^{(t)}$. ■

*Lemma A.7:* For all $i \in [n]$ and $t \in \{0, 1, \ldots, T-1\}$, the expcted suare norm of the memory vector $\boldsymbol{m}_i^{(t)}$ is bounded by

$$\mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 \leq 2\eta^2 \Lambda Q^2 \left( (1+\alpha L_G)^2 + \frac{\alpha^2 \sigma_H^2}{m_B} \right) G^2,$$

where $\Lambda = \frac{(1-\lambda)(1+\frac{1}{c})}{1-(1-\lambda)(1+c)}$ and $0 < c < \frac{\lambda}{1-\lambda}$ with $\lambda = \frac{k}{d}$.

*Proof:*

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{m}_i^{(t+1)}\|^2 \\
&= \mathbb{E}\|\boldsymbol{m}_i^{(t)} + \Delta_i^{(t)} - \boldsymbol{g}_i^{(t)}\|^2 \\
&\leq (1-\lambda)\mathbb{E}\|\boldsymbol{m}_i^{(t)} + \Delta_i^{(t)}\|^2 \\
&\overset{(a)}{\leq} (1-\lambda)(1+c)\mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 + (1-\lambda)(1+\frac{1}{c})\mathbb{E}\|\boldsymbol{\Delta}_i^{(t)}\|^2 \\
&\overset{(b)}{\leq} (1-\lambda)(1+c)\mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 + (1-\lambda)(1+\frac{1}{c})\eta^2 Q \sum_{q=0}^{Q-1} \mathbb{E}\|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\|^2, \\
&\overset{(c)}{\leq} (1-\lambda)(1+c)\mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 + (1-\lambda)(1+\frac{1}{c}) \underbrace{\eta^2 Q^2 \left( 2(1+\alpha L_G)^2 + 2\frac{\alpha^2\sigma_H^2}{m_B} \right) G^2}_{\triangleq \tilde{G}^2} \\
&\leq (1-\lambda)(1+\frac{1}{c})\tilde{G}^2 \sum_{j=0}^{\infty} [(1-\lambda)(1+c)]^j. \qquad (40)
\end{aligned}$$

where (a) is followed by Lemma A.1 for any $\tau > 0$, (b) is followed by Jensen's inequality and (c) is followed by Lemma A.5. Let $0 < c < \frac{\lambda}{1-\lambda}$ such that $(1-\lambda)(1+c) < 1$ and

$$\sum_{j=0}^{\infty} [(1-\lambda)(1+c)]^j = \frac{1}{1-(1-\lambda)(1+c)}, \qquad (41)$$

which yields the result by defining $\Lambda = \frac{(1-\lambda)(1+\frac{1}{c})}{1-(1-\lambda)(1+c)}$. ■

*Lemma A.8:* Considering the power constraint $(1/M)\mathbb{E}\|(\sqrt{\rho^{(t)}}/\eta)\boldsymbol{g}_i^{(t)}\|^2 \leq P_i$, we have, for $t \in \{0, 1, \ldots, T-1\}$,

$$\frac{1}{\rho^{(t)}} \leq \frac{4Q^2G^2(\Lambda+1)}{MP_{\min}} \left( (1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B} \right).$$

*Proof:* Note that $\mathbb{E}\|\frac{\sqrt{\rho^{(t)}}}{\eta}\boldsymbol{g}_i^{(t)}\|^2 = \frac{\rho^{(t)}}{\eta^2}\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2 \leq MP_i$. Define

$$\rho = \min_t \min_i \frac{\eta^2 MP_i}{\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2}$$

such that $\rho_t > \rho$. Define $P_{\min} = \min_{i\in[n]} P_i$. Then, we have

$$\begin{aligned}
\frac{1}{\rho} &= \max_t \max_i \frac{\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2}{\eta^2 M P_i} \\
&\overset{(a)}{\leq} \max_t \max_i \frac{\mathbb{E}\|\boldsymbol{m}_i^{(t)} + \Delta_i^{(t)}\|^2}{\eta^2 M P_i} \\
&\leq \frac{2}{\eta^2 M P_{\min}} \max_{t,i} \left( \mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_i^{(t)}\|^2 \right) \\
&\leq \frac{2}{\eta^2 M P_{\min}} \max_{t,i} \left( \mathbb{E}\|\boldsymbol{m}_i^{(t)}\|^2 + Q\eta^2 \sum_{q=0}^{Q-1} \mathbb{E}\|\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)})\|^2 \right) \\
&\overset{(b)}{\leq} \frac{4Q^2G^2(\Lambda+1)}{M P_{\min}} \left( (1+\alpha L_G)^2 + \frac{\alpha^2 \sigma_H^2}{m_B} \right),
\end{aligned} \tag{42}$$

where (a) is due to the fact that $\|\text{Comp}_k(\boldsymbol{x})\|^2 \leq \|\boldsymbol{x}\|^2$, and (b) is followed by Lemma A.7 and A.5. The above inequality yields the desired result. ■

### *B. Proofs of Theorem 4.1 and Theorem 4.2*

The proof of Theorem 4.1 is based on the perturbed iterate analysis as in [45]. To this end, we define the maintained virtual sequence $\{\hat{\boldsymbol{\theta}}^{(t)}\}_{t=0,\ldots,T}$ as follows:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \boldsymbol{\Delta}_i^{(t)} - \left( \frac{\eta^{(t)}}{\mu_h rn \sqrt{\rho^{(t)}}} \boldsymbol{n}_{\text{est}}^{(t)} + \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \left( \frac{|h_i^{(t)}|}{\mu_h} - 1 \right) \boldsymbol{g}_i^{(t)} \right), \tag{43}$$

where $\hat{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}^{(0)}$. The relation between the true sequence $\{\boldsymbol{\theta}^{(t)}\}_{t=0,\ldots,T}$ and the virtual sequence $\{\hat{\boldsymbol{\theta}}^{(t)}\}_{t=0,\ldots,T}$ is given by the following Lemma A.9.

*Lemma A.9:* The true sequence $\{\boldsymbol{\theta}^{(t)}\}_{t=0,\ldots,T}$ and the virtual sequence $\{\tilde{\boldsymbol{\theta}}^{(t)}\}_{t=0,\ldots,T}$ satisfy the following equality: $\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)} = (1/rn)\sum_{i=1}^n \boldsymbol{m}_i^{(t)}$, where the expectation is over the randomness of channel gain and device selection.

*Proof:* Following the Assumption 5 and the global update (17), we have

$$\begin{aligned}
&\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}}^{(t+1)} \\
&= \boldsymbol{\theta}^{(t)} - \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \frac{|h_t^{(i)}|}{\mu_h} \boldsymbol{g}_i^{(t)} - \frac{\eta^{(t)}}{\mu_h rn \sqrt{\rho^{(t)}}} \boldsymbol{n}_{\text{est}}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \boldsymbol{\Delta}_i^{(t)} + \frac{\eta^{(t)}}{\mu_h rn \sqrt{\rho^{(t)}}} \boldsymbol{n}_{\text{est}}^{(t)} \\
&\quad + \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \left( \frac{|h_i^{(t)}|}{\mu_h} - 1 \right) \boldsymbol{g}_i^{(t)} \\
&\overset{(a)}{=} \boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} (\boldsymbol{\Delta}_i^{(t)} - \boldsymbol{g}_i^{(t)})
\end{aligned}$$

$$\overset{(b)}{=} \boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}(\boldsymbol{m}_i^{(t+1)} - \boldsymbol{m}_i^{(t)})$$

$$\overset{(c)}{=} \boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{rn}\sum_{i=1}^{n}(\boldsymbol{m}_i^{(t+1)} - \boldsymbol{m}_i^{(t)})$$

$$= \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=0}^{t}(\boldsymbol{m}_i^{(j+1)} - \boldsymbol{m}_i^{(j)})$$

$$\overset{(d)}{=} \frac{1}{rn}\sum_{i=1}^{n}\boldsymbol{m}_i^{(t+1)}, \tag{44}$$

where (a) is followed by $\frac{|h_t^{(i)}|}{\mu_h}\boldsymbol{g}_i^{(t)} = \boldsymbol{g}_i^{(t)} + (\frac{|h_t^{(i)}|}{\mu_h} - 1)\boldsymbol{g}_i^{(t)}$; (b) is followed by memory update rule $\boldsymbol{m}_i^{(t+1)} = \boldsymbol{m}_i^{(t)} + \left(\boldsymbol{\Delta}_i^{(t)} - \boldsymbol{g}_i^{(t)}\right)$; (c) is followed by the fact that if $i \notin \mathcal{I}^{(t)}$, the memory will not be updated, i.e., $\boldsymbol{m}_i^{(t+1)} - \boldsymbol{m}_i^{(t)} = \mathbf{0}$ for $i \notin \mathcal{I}^{(t)}$; (d) is directly obtained by telescope sum. ■

Recording the definition of global loss function $F(\boldsymbol{\theta})$ and the global update rule, we consider the perturbed sequence and begin the derivation by Lemma A.2, i.e., for any $t = 0, 1, \ldots, T-1$, we have

$$\mathbb{E}_n\left[F(\hat{\boldsymbol{\theta}}^{(t+1)})\right] \leq F(\hat{\boldsymbol{\theta}}^{(t)}) - \left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h} - 1\right)\boldsymbol{g}_i^{(t)}\right\rangle$$
$$+ \frac{L_F}{2}\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \left(\frac{\eta^{(t)}}{\mu_h rn\sqrt{\rho^{(t)}}}\boldsymbol{n}_{\text{est}}^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h} - 1\right)\boldsymbol{g}_i^{(t)}\right)\right\|^2 \tag{45}$$

Taking expectation on (45) over the estimate error $\boldsymbol{n}_{\text{est}}^{(t)}$ and following Assumption 5 yield

$$\mathbb{E}_n\left[F(\hat{\boldsymbol{\theta}}^{(t+1)})\right] \leq F(\hat{\boldsymbol{\theta}}^{(t)}) - \left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h} - 1\right)\boldsymbol{g}_i^{(t)}\right\rangle$$
$$+ \frac{L_F}{2}\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h} - 1\right)\boldsymbol{g}_i^{(t)}\right\|^2 + \frac{L_F\eta^2}{2\rho^{(t)}}\frac{dv^{(t)}}{\mu_h^2 r^2 n^2} \tag{46}$$

In the following, our target is to bound the inner-product term and the square norm term. First, we consider the inner-product term and take expectation over SGD, device sampling, random sparsification, and channel fading conditioned on time $t$, yielding

$$-\mathbb{E}\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h} - 1\right)\boldsymbol{g}_i^{(t)}\right\rangle$$
$$\overset{(a)}{=} -\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{\eta}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)})\right] + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mu_h}{\mu_h} - 1\right)\mathbb{E}\left[\boldsymbol{g}_i^{(t)}\right]\right\rangle$$

$$
\begin{aligned}
&= -\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{\eta}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)})\right]\right\rangle \\
&= -\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{\eta}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - Q\nabla F_i(\hat{\boldsymbol{\theta}}^{(t)}) + Q\nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right]\right\rangle \\
&\overset{(b)}{=} -\eta Q\|\nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2 - \eta\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - Q\nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right]\right\rangle \\
&\overset{(c)}{\leq} -\frac{\eta Q}{2}\|\nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2 + \frac{\eta}{2Q}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - Q\nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right]\right\|^2
\end{aligned} \tag{47}
$$

where (a) is followed by tower rule and Assumption 6; (b) is due to the definition of $F(\boldsymbol{\theta}) \triangleq 1/n\sum_{i=1}^{n}F_i(\boldsymbol{\theta})$, and (c) is followed by Young's inequality (c.f. Lemma A.1). Then we bound the last term in (47) as follows:

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - Q\nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right]\right\|^2 \\
&\overset{(a)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbb{E}\left[\sum_{q=0}^{Q-1}\left(\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) + \nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)}) + \nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right)\right]\right\|^2 \\
&\overset{(b)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\left(3\left\|\mathbb{E}\left[\sum_{q=0}^{Q-1}\left(\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)})\right)\right]\right\|^2\right. \\
&\qquad + 3\left\|\mathbb{E}\left[\sum_{q=0}^{Q-1}\left(\nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)})\right)\right]\right\|^2 \\
&\qquad \left. + 3\left\|\mathbb{E}\left[\sum_{q=0}^{Q-1}\left(\nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F_i(\hat{\boldsymbol{\theta}}^{(t)})\right)\right]\right\|^2\right) \\
&\overset{(c)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{12\alpha^2Q^2L^2\sigma_G^2}{m_B} + 3L_F^2Q\sum_{q=0}^{Q-1}\mathbb{E}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_i^{(t,q)}\|^2 + 3L_F^2Q^2\mathbb{E}\|\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)}\|^2\right)
\end{aligned} \tag{48}
$$

where (a) and (b) are followed by Jensen's inequality and (c) is followed by Lemma A.2, Lemma A.3, and Jensen's inequality. The relation between $\|\nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2$ and $\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2$ is given by

$$
\begin{aligned}
\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 &\leq 2\|\nabla F(\boldsymbol{\theta}^{(t)}) - \nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2 + 2\|\nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2 \\
&\leq 2L_F^2\|\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)}\|^2 + 2\|\nabla F(\hat{\boldsymbol{\theta}}^{(t)})\|^2.
\end{aligned} \tag{49}
$$

Combining (47) with (48) and (49) yields

$$
\begin{aligned}
&-\left\langle \nabla F(\hat{\boldsymbol{\theta}}^{(t)}), \frac{\eta}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{q=0}^{Q-1}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)})\right]\right\rangle \\
\leq& -\frac{\eta Q}{4}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 + \frac{6\eta\alpha^2 Q L^2\sigma_G^2}{m_B} + \frac{3}{2}\eta L_F^2\frac{1}{n}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}_i^{(t,q)}\|^2 + 2\eta Q L_F^2\mathbb{E}\|\boldsymbol{\theta}^{(t)}-\hat{\boldsymbol{\theta}}^{(t)}\|^2 \\
\overset{(a)}{=}& -\frac{\eta Q}{4}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 + \frac{6\eta\alpha^2 Q L^2\sigma_G^2}{m_B} + \frac{3}{2}\eta L_F^2\frac{1}{n}\sum_{i=1}^{n}\sum_{q=0}^{Q-1}\mathbb{E}\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}_i^{(t,q)}\|^2 + 2\eta Q L_F^2\mathbb{E}\left\|\frac{1}{rn}\sum_{i=1}^{n}\boldsymbol{m}_i^{(t)}\right\|^2 \\
\overset{(b)}{\leq}& -\frac{\eta Q}{4}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 + \frac{6\eta\alpha^2 Q L^2\sigma_G^2}{m_B} + \frac{4\eta^3 Q^3 L_F^2\Lambda}{r^2}\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)G^2 \\
&+ \frac{3}{2}\eta^3 Q^3 L_F^2(40(\sigma_F^2+\gamma_F^2) + 40\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2),
\end{aligned} \tag{50}
$$

where (a) is followed by Lemma A.9 and (b) is followed by Jensen inequality, Lemma A.7 and Lemma A.6.

Then, we consider the square norm term in (46) by taking expectation over SGD, device sampling, random sparsification, and channel fading conditioned on time $t$, i.e.,

$$
\begin{aligned}
&\frac{L_F}{2}\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)} + \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)\boldsymbol{g}_i^{(t)}\right\|^2 \\
\leq& L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)}\right\|^2 + L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)\boldsymbol{g}_i^{(t)}\right\|^2
\end{aligned} \tag{51}
$$

Then we bound the first term in (51) as follow:

$$
\begin{aligned}
&L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)}\right\|^2 \\
=&\eta^2 L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\sum_{q=0}^{Q-1}\left(\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)})\right)\right\|^2 \\
\overset{(a)}{\leq}&\eta^2 Q L_F\sum_{q=0}^{Q-1}\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)})\right\|^2 \\
=&\eta^2 Q L_F\sum_{q=0}^{Q-1}\mathbb{E}\Bigg\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\Big(\hat{\nabla}F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) + \nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)}) \\
&+ \nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)}) + \nabla F(\boldsymbol{\theta}^{(t)})\Big)\Bigg\|^2
\end{aligned}
$$

$$\leq \eta^2 Q L_F \sum_{q=0}^{Q-1} \Bigg( 4\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)})\big)\Bigg\|^2 + 4\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \nabla F(\boldsymbol{\theta}^{(t)})\Bigg\|^2$$
$$+ 4\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)})\big)\Bigg\|^2 + 4\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\big)\Bigg\|^2 \Bigg) \tag{52}$$

where (a) is due to the sum is changeable and followed by Jensen's inequality. We bound the first three square $l_2$-norm terms as follows.

$$\begin{aligned}
&\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)})\big)\Bigg\|^2 \\
\overset{(a)}{\leq} &\mathbb{E}\left[ \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \Bigg\| \big(\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)})\big)\Bigg\|^2 \right] \\
\overset{(b)}{=} &\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Bigg\| \big(\hat{\nabla} F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}_i^{(t,q)})\big)\Bigg\|^2 \\
\overset{(c)}{\leq} &\sigma_F^2,
\end{aligned} \tag{53}$$

where (a), (b), and (c) are followed by Jensen's inequality, tower rule and Lemma A.3, respectively. Similarly,

$$\begin{aligned}
&\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)})\big)\Bigg\|^2 \\
\leq &\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Bigg\| \nabla F_i(\boldsymbol{\theta}_i^{(t,q)}) - \nabla F_i(\boldsymbol{\theta}^{(t)})\Bigg\|^2 \\
\overset{(a)}{\leq} &\frac{L_F^2}{n} \sum_{i=1}^{n} \mathbb{E}\Bigg\| \boldsymbol{\theta}_i^{(t,q)} - \boldsymbol{\theta}^{(t)}\Bigg\|^2 \\
\overset{(b)}{\leq} &L_F^2 (40Q^2\eta^2(\sigma_F^2 + \gamma_F^2) + 40Q^2\eta^2 \mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2)
\end{aligned} \tag{54}$$

where (a) and (b) are followed Lemma A.2 and A.6. For the third $l_2$-norm term, we bound it as follows:

$$\begin{aligned}
&\mathbb{E}\Bigg\| \frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \big(\nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\big)\Bigg\|^2 \\
\leq &\frac{1}{rn} \sum_{i\in\mathcal{I}^{(t)}} \mathbb{E}\Bigg\| \big(\nabla F_i(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\big)\Bigg\|^2 \\
\overset{(a)}{\leq} &\gamma_F^2
\end{aligned} \tag{55}$$

where (a) follows Lemma A.4. Therefore, we have the following bound:

$$
\begin{aligned}
&L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\boldsymbol{\Delta}_i^{(t)}\right\|^2\\
\leq&4\eta^2Q^2L_F\left(\sigma_F^2+L_F^2(40Q^2\eta^2(\sigma_F^2+\gamma_F^2)+40Q^2\eta^2\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2)+\gamma_F^2+\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\right)
\end{aligned}
\tag{56}
$$

The second term in (51) can be bounded as

$$
\begin{aligned}
&L_F\mathbb{E}\left\|\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)\boldsymbol{g}_i^{(t)}\right\|^2\\
\overset{(a)}{\leq}&L_F\mathbb{E}\left[\frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\left\|\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)\boldsymbol{g}_i^{(t)}\right\|^2\right]\\
\overset{(b)}{=}&L_F\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)\boldsymbol{g}_i^{(t)}\right\|^2\\
\overset{(c)}{\leq}&4\eta^2L_FQ^2G^2(\Lambda+1)\left((1+\alpha L_G)^2+\frac{\alpha^2\sigma_H^2}{m_B}\right)\mathbb{E}\left[\left(\frac{|h_i^{(t)}|}{\mu_h}-1\right)^2\right]\\
=&4\eta^2L_FQ^2G^2(\Lambda+1)\left((1+\alpha L_G)^2+\frac{\alpha^2\sigma_H^2}{m_B}\right)\left(\frac{\sigma_h^2}{\mu_h^2}-1\right),
\end{aligned}
\tag{57}
$$

where (a), (b) and (c) are followed by Jensen inequality, tower rule and the bound of $\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2$ obtained by the same technique used in the proof of Lemma A.8, respectively.

Substituting (50), (56) and (57) into (46) by taking expectation, we have the following per-round convergence bound:

$$
\begin{aligned}
\mathbb{E}_n\left[F(\hat{\boldsymbol{\theta}}^{(t+1)})\right]\leq F(\hat{\boldsymbol{\theta}}^{(t)})-\frac{\eta Q}{4}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2+\frac{6\eta\alpha^2QL^2\sigma_G^2}{m_B}+\frac{4\eta^3Q^3L_F^2\Lambda}{r^2}\left((1+\alpha L_G)^2+\frac{\alpha^2\sigma_H^2}{m_B}\right)G^2\\
+\frac{3}{2}\eta^3Q^3L_F^2\left(40(\sigma_F^2+\gamma_F^2)+40\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\right)\\
+4\eta^2Q^2L_F\left(\sigma_F^2+L_F^2(40Q^2\eta^2(\sigma_F^2+\gamma_F^2)+40Q^2\eta^2\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2)+\gamma_F^2+\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2\right)\\
+4\eta^2L_FQ^2G^2(\Lambda+1)\left((1+\alpha L_G)^2+\frac{\alpha^2\sigma_H^2}{m_B}\right)\left(\frac{\sigma_h^2}{\mu_h^2}-1\right)+\frac{L_F\eta^2}{2\rho^{(t)}}\frac{dv^{(t)}}{r^2n^2}
\end{aligned}
$$

Let $\eta$ satisfy the following inequality

$$
60\eta^2Q^2L_F^2+160\eta^3Q^3L_F^3+4\eta QL_F\leq\frac{1}{8},
\tag{58}
$$

which simplifies the inequality as follows

$$\mathbb{E}_n\left[F(\hat{\boldsymbol{\theta}}^{(t+1)})\right] \leq F(\hat{\boldsymbol{\theta}}^{(t)}) - \frac{\eta Q}{4}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 + \frac{6\eta\alpha^2 Q L^2 \sigma_G^2}{m_B} + \frac{4\eta^3 Q^3 L_F^2 \Lambda}{r^2}\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)G^2$$
$$+ 60\eta^3 Q^3 L_F^2(\sigma_F^2+\gamma_F^2) + 4\eta^2 Q^2 L_F\left(\sigma_F^2 + L_F^2\left(40Q^2\eta^2(\sigma_F^2+\gamma_F^2)\right) + \gamma_F^2\right)$$
$$+ 4\eta^2 L_F Q^2 G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\left(\frac{\sigma_h^2}{\mu_h^2} - 1\right) + \frac{L_F\eta^2}{2\rho^{(t)}}\frac{dv^{(t)}}{r^2 n^2} \quad (59)$$

By Lemma A.8, rearranging, taking telescope sum and taking expectation over all randomness, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^{(t)})\|^2 \leq \frac{8}{\eta Q T}\left(F(\boldsymbol{\theta}^{(0)}) - F^*\right) + \frac{48\alpha^2 L^2\sigma_G^2}{m_B}$$
$$+ \frac{32\eta^2 Q^2 L_F^2\Lambda}{r^2}\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)G^2 + 480\eta^2 Q^2 L_F^2(\sigma_F^2+\gamma_F^2)$$
$$+ 32\eta Q L_F\left(\sigma_F^2 + L_F^2\left(40Q^2\eta^2(\sigma_F^2+\gamma_F^2)\right) + \gamma_F^2\right)$$
$$+ 32\eta L_F Q G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\left(\frac{\sigma_h^2}{\mu_h^2} - 1\right)$$
$$+ \frac{16\eta L_F d Q G^2(\Lambda+1)}{r^2 n^2 M P_{\min}}\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\frac{1}{T}\sum_{t=0}^{T-1} v^{(t)}, \quad (60)$$

where we use $\hat{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}^{(0)}$, $F^* = \min_{\boldsymbol{\theta}\in\mathbb{R}^d} F(\boldsymbol{\theta})$, and $\eta \leq 1/L_F$ to obtain the desired result of Theorem 4.1.

The proof of Theorem 4.2 is analogous to that of Theorem 4.1 in principal. It is worth noting that we should carefully design the parameters of $\alpha^{(t)}$ and $\eta^{(t)}$ such that we can arrive at a first-order regression inequality similar to (59). In addition, we modify the memory bound for adaptive learning rate derived in [53, Lemma 4]. Next, we take a telescopic sum, and use the inequalities $\sum_{t=0}^{T-1}\eta^{(t)} \geq \xi\ln\left((T+a-1)/a\right)$, $\sum_{t=0}^{T-1}(\eta^{(t)})^2 \leq \xi^2/(a-1)$ and $\sum_{t=0}^{T-1}(\eta^{(t)})^3 \leq \xi^3/2(a-1)^2$, which completes the proof.

### C. Proofs of Theorem 5.1

The effective global update of Air-meta-pFL is given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \frac{1}{rn}\sum_{i\in\mathcal{I}^{(t)}}\frac{|h_i^{(t)}|}{\mu_h}\boldsymbol{g}_i^{(t)} - \frac{\eta^{(t)}}{\mu_h r n\sqrt{\rho^{(t)}}}\boldsymbol{n}_{\text{est}}^{(t)} \quad (61)$$

which is a noisy weighted-sum SGD update. We first define the Markov chain as follows

$$\mathcal{D}_{1:n} \to \mathcal{D}_{\mathcal{I}_{[T]}} \to \mathcal{B}_{\mathcal{I}_{[T]}} \to \boldsymbol{\theta}_{[T]} \to \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is short for the output of Air-meta-pFL, i.e., $\boldsymbol{\theta}^{(T)}$, for the ease of exhibition; $\mathcal{D}_{\mathcal{I}_{[T]}} = \{\mathcal{D}_{\mathcal{I}_{(0)}}, \mathcal{D}_{\mathcal{I}_{(1)}}, \ldots, \mathcal{D}_{\mathcal{I}_{(T-1)}}\}$ with $\mathcal{D}_{\mathcal{I}_{(t)}} = \{\mathcal{D}_i\}_{i \in \mathcal{I}^{(t)}}$; $\mathcal{B}_{\mathcal{I}_{[T]}}$ is defined as a similar way; and $\boldsymbol{\theta}_{[T]} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}\}$. Based on this Markov Chain, we have the following inequalities by data processing inequality:

$$I(\boldsymbol{\theta}; \mathcal{D}_{1:n}) \leq I(\boldsymbol{\theta}_{[T]}; \mathcal{D}_{1:n}) \leq I(\boldsymbol{\theta}_{[T]}; \mathcal{D}_{\mathcal{I}_{[T]}}) \leq I(\boldsymbol{\theta}_{[T]}; \mathcal{B}_{\mathcal{I}_{[T]}}).$$

We calculate the term $I(\boldsymbol{\theta}_{[T]}; \mathcal{B}_{\mathcal{I}_{[T]}})$ as follows:

$$\begin{aligned} I(\boldsymbol{\theta}_{[T]}; \mathcal{B}_{\mathcal{I}_{[T]}}) &= \sum_{t=1}^{T} I(\boldsymbol{\theta}^{(t)}; \mathcal{B}_{\mathcal{I}_{[T]}} \mid \boldsymbol{\theta}_{[t-1]}) \\ &\overset{(a)}{=} \sum_{t=1}^{T} I(\boldsymbol{\theta}^{(t)}; \mathcal{B}_{\mathcal{I}^{(t)}} \mid \boldsymbol{\theta}^{(t-1)}) \\ &= \sum_{t=1}^{T} \left( h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}) - h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathcal{B}_{\mathcal{I}^{(t)}}) \right), \end{aligned} \tag{62}$$

where $h(x|y)$ is the entropy of random variable $x$ conditioned on $y$, the first equality is followed by the chain rule of mutual information, (a) is followed by the update rule in (61) and the assumption that the sampling strategy is independent of the parameters and the previous samplings. In order to bound the entropy $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)})$, note that conditioned on $\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)}$, we have $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)}) = h(\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)} \mid \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)})$. We next bound the second moment

of the random variable $\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)}$ as below.

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)}\|^2 \\
=&\mathbb{E}\left\|\frac{1}{\mu_h rn}\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|\boldsymbol{g}_i^{(t-1)}\right\|^2 + \frac{d\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}} \\
\overset{(a)}{\leq}&\frac{1}{\mu_h^2 rn}\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2\mathbb{E}\left\|\boldsymbol{g}_i^{(t-1)}\right\|^2 + \frac{d\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}} \\
\overset{(b)}{\leq}&\frac{1}{\mu_h^2 rn}\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2\mathbb{E}\left\|\boldsymbol{m}_i^{(t-1)} + \boldsymbol{\Delta}_i^{(t)}\right\|^2 + \frac{d\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}} \\
\leq&\frac{1}{\mu_h^2 rn}\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2\left(2\mathbb{E}\left\|\boldsymbol{m}_i^{(t-1)}\right\|^2 + 2\mathbb{E}\left\|\boldsymbol{\Delta}_i^{(t)}\right\|^2\right) + \frac{d\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}} \\
\overset{(c)}{\leq}&\underbrace{\frac{1}{\mu_h^2 rn}\left(4\eta^2 Q^2 G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\right)\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2}_{C_1} + \underbrace{\frac{d\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}}}_{C_2},
\end{aligned}
\tag{63}
$$

where (a) follows the convexity of $\|\cdot\|^2$, (b) is by the definition of $\boldsymbol{g}_i^{(t)}$, and (c) is as a result of Lemma A.5 and Lemma A.7. Since $1/\rho^{(t)} < \infty$ by Lemma A.8, $\mathbb{E}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\vartheta}^{(t-1)}\|^2$ is bounded, which leads to $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)})$ upper-bounded by the entropy of the Gaussian distribution with zero mean and variance $\sqrt{(C_1+C_2)/d}\boldsymbol{I}_d$, i.e., $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\vartheta}^{(t-1)}) \leq \frac{d}{2}\log\left(\frac{2\pi e(C_1+C_2)}{d}\right)$. Since this bound holds for all values of $\boldsymbol{\vartheta}^{(t-1)}$, we conclude that $h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}) \leq \frac{d}{2}\log\left(\frac{2\pi e(C_1+C_2)}{d}\right)$. Combining with

$$
\begin{aligned}
h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathcal{B}_{\mathcal{I}^{(t-1)}}) =& h\left(\frac{\eta}{\mu_h rn\sqrt{\rho^{(t-1)}}}\boldsymbol{n}_{\text{est}}^{(t-1)}\right) \\
=&\frac{d}{2}\log\left(2\pi e\frac{\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}}\right),
\end{aligned}
$$

we have

$$
\begin{aligned}
&h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}) - h(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathcal{B}_{\mathcal{I}^{(t)}}) \\
\leq& \frac{d}{2}\log\left(\frac{2\pi e(C_1+C_2)}{d}\right) - \frac{d}{2}\log\left(2\pi e\frac{\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}}\right) \\
=& \frac{d}{2}\log\left(1 + \frac{2\pi e\frac{1}{\mu_h^2 rn}\left(4\eta^2 Q^2 G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\right)\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2}{2\pi e\frac{\eta^2 v^{(t)}}{\mu_h^2 r^2 n^2 \rho^{(t-1)}}}\right) \\
\leq& \frac{d}{2}\log\left(1 + \frac{MP_{\max}\left(4Q^2 G^2(\Lambda+1)\left((1+\alpha L_G)^2 + \frac{\alpha^2\sigma_H^2}{m_B}\right)\right) rn\sum_{i\in\mathcal{I}^{(t-1)}}|h_i^{(t-1)}|^2}{dv^{(t)}\epsilon_g}\right),
\end{aligned}
\tag{64}
$$

TABLE III

EXPERIMENTAL PARAMETERS FOR CIFAR10 AND FASHION-MNIST

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| Mini-batch size | $m_B = 64$ | Data heterogeneity | 2 classes of data for each device |
| Local SGD | $Q = 5$ | Number of devices | $n = 20$ |
| Fraction of active device | $r = 0.5$ | Number of rounds | $T = 100$ |
| Meta-learning rate | $\eta = 0.05$ | Inner learning rate | $\alpha = 0.05$ |
| Model | CNN | Monte Carlo | 10 |

where we use $\rho^{(t)} \leq MP_{\max}/\epsilon_g$ with the assumption of $\mathbb{E}\|\boldsymbol{g}_i^{(t)}\|^2 \geq \eta^2\epsilon_g$ and define $P_{\max} = \max_{i\in[n]} P_i$. Therefore,

$$|\Delta_\tau| \leq \sqrt{\frac{d\sigma^2}{n}\sum_{t=0}^{T-1}\log\left(1+\frac{MP_{\max}rnC_g\sum_{i\in\mathcal{I}^{(t)}}|h_i^{(t)}|^2}{dv^{(t)}\epsilon_g}\right)}, \tag{65}$$

which completes the proof by defining $C_g = 4Q^2G^2(\Lambda+1)((1+\alpha L_G)^2+\alpha^2\sigma_H^2/m_B)$.

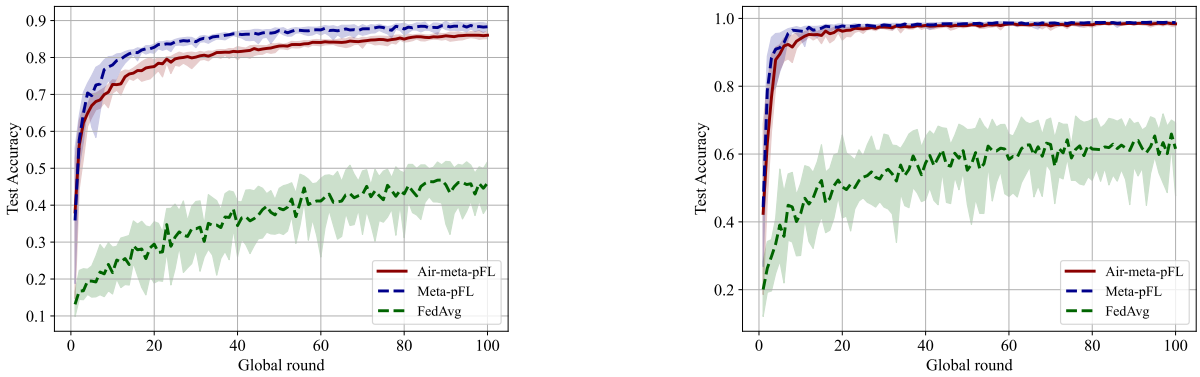
### *D. Additional Experimental Results*

To validate the generalization capacity of the proposed Air-meta-pFL, we conduct additional experiments on CIFAR10 [54] and Fashion-MNIST [55] data sets with the experimental parameters specified in Table III. The task of each device is to distinguish among $N = 2$ different classes, which are pre-assigned to each device from 10 total classes, with $K = 32$ shots for each class (see Fig. 3). Other parameters that are not listed here remain the same as in Sec. VI-A. We use a simple CNN network that consists of two convolutional layers with ReLU activation and max pooling, followed by two fully connected layers for 10-class classification.

Fig. 10(a) and Fig. 10(b) show the test accuracy versus global rounds $T$ with Air-meta-pFL and other benchmarks on CIFAR10 [R7] and Fashion-MNIST [R8], respectively. This result shows that the proposed Air-meta-pFL achieves comparable performance with vanilla meta-pFL [15], and significantly outperforms FedAvg [8] across different data sets, demonstrating its good generalization capabilities.

## REFERENCES

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.




(a) (b)

Fig. 10. Test accuracy versus global rounds $T$ on data set (a) CIFAR10; and (b) Fashion-MNIST.


[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[5] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[6] H. Sun, H. Tian, W. Ni, J. Zheng, D. Niyato, and P. Zhang, "Federated low-rank adaptation for large models fine-tuning over wireless networks," *IEEE Trans Wireless Commun.*, vol. 24, no. 1, pp. 659–675, 2025.

[7] W. Ni, H. Sun, H. Ao, and H. Tian, "Federated intelligence: When large ai models meet federated fine-tuning and collaborative reasoning at the network edge," *IEEE Internet Things Mag.*, pp. 1–8, 2025.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artificial Intelligence and Statistics*, FL, USA, Apr. 2017.

[9] L. Chen, S. T. Jose, I. Nikoloska, S. Park, T. Chen, O. Simeone *et al.*, "Learning with limited samples: Meta-learning and applications to communication systems," *Foundations and Trends® in Signal Processing*, vol. 17, no. 2, pp. 79–208, 2023.

[10] O. Simeone, *Machine learning for engineers*. Cambridge University Press, 2022.

[11] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[12] M. Setayesh, X. Li, and V. W. Wong, "Perfedmask: Personalized federated learning with optimized masking vectors," in *Proc. International Conference on Learning Representations*, May, Kigali Rwanda 2023.

[13] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[14] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–17, 2022.

[15] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic

meta-learning approach," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2020.

[16] B. Wang, Z. Yuan, Y. Ying, and T. Yang, "Memory-based optimization methods for model-agnostic meta-learning and personalized federated learning." *J. Mach. Learn. Res.*, vol. 24, pp. 145–1, 2023.

[17] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar, "Adaptive gradient-based meta-learning methods," in *Proc. Advances in Neural Information Processing Systems*, BC, Canada, Dec. 2019.

[18] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.

[19] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[20] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[21] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 302–323, 2019.

[22] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proc. International conference on machine learning*, Washington, USA, Jun. 2011.

[23] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2020.

[24] ——, "Wireless federated langevin monte carlo: Repurposing channel noise for bayesian sampling and privacy," *IEEE Trans. Wireless Commun.*, 2022.

[25] H. H. Yang, Z. Chen, T. Q. Quek, and H. V. Poor, "Revisiting analog over-the-air machine learning: The blessing and curse of interference," *IEEE J Sel. Top. Signal Process.*, vol. 16, no. 3, pp. 406–419, 2021.

[26] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, 2020.

[27] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, 2021.

[28] H. Wen, H. Xing, and O. Simeone, "Convergence analysis of over-the-air FL with compression and power control via clipping," in *Proc. IEEE Global Communications Conference*, Kuala Lumpur, Malaysia, Dec. 2023.

[29] ——, "AirFL-Mem: Improving communication-learning trade-off by long-term memory," *Accepted by IEEE Wireless Communications and Networking Conference (WCNC) 2024*, 2024.

[30] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[31] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, 2021.

[32] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[33] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. International conference on machine learning*, Sydney, Australia, Aug. 2017.

[34] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, "Meta-learning with implicit gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[35] S. Yue, J. Ren, J. Xin, D. Zhang, Y. Zhang, and W. Zhuang, "Efficient federated meta-learning over multi-access wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1556–1570, 2022.

[36] C. You, K. Guo, G. Feng, P. Yang, and T. Q. Quek, "Automated federated learning in mobile edge networks—fast adaptation and convergence," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13 571–13 586, 2023.

[37] Y. Li, X. Qin, H. Chen, K. Han, and P. Zhang, "Energy-aware edge association for cluster-based personalized federated learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6756–6761, 2022.

[38] Z. Zhao, J. Wang, W. Hong, T. Q. Quek, Z. Ding, and M. Peng, "Ensemble federated learning with non-iid data in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3557–3571, 2024.

[39] Z. Chen, W. Yi, Y. Liu, and A. Nallanathan, "Knowledge-aided federated learning for energy-limited wireless networks," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3368–3386, 2023.

[40] Z. Chen, Z. Li, H. H. Yang, and T. Q. Quek, "Personalizing federated learning with over-the-air computations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, Jun 2023.

[41] M. Mortaheb, C. Vahapoglu, and S. Ulukus, "Personalized federated multi-task learning over wireless fading channels," *Algorithms*, vol. 15, no. 11, p. 421, 2022.

[42] H. U. Sami and B. Güler, "Over-the-air personalized federated learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, May 2022.

[43] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, vol. 23, no. 1, p. 126, 2021.

[44] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2021.

[45] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Advances in Neural Information Processing Systems*, Montréal Canada, Dec. 2018.

[46] N. K. Jha, H. Guo, and V. K. Lau, "Analog product coding for over-the-air aggregation over burst-sparse interference multiple-access channels," *IEEE Trans. Signal Process.*, vol. 72, pp. 157–172, 2023.

[47] H. Xing, O. Simeone, and S. Bi, "Federated learning over wireless device-to-device networks: Algorithms and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3723–3741, 2021.

[48] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[49] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," in *Proc. National Academy of Science*, vol. 106, no. 45, 2009, pp. 18 914–18 919.

[50] O. Rivasplata, "Subgaussian random variables: An expository note," *Internet publication, PDF*, vol. 5, 2012.

[51] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proc. the annual meeting of the cognitive science society*, Massachusetts, USA, Jul. 2011.

[52] R. R. T., "Over-the-air federated learning under time-varying wireless channels using OTFS," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 1782–1787, 2025.

[53] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019.

[54] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[55] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1708.07747