# Over-the-Air Federated Learning with Joint Adaptive Computation and Power Control

Haibo Yang, Peiwen Qiu, Jia Liu and Aylin Yener
Dept. of Electrical and Computer Engineering, The Ohio State University
{yang.9292, qiu.617}@osu.edu, liu@ece.osu.edu, yener@ece.osu.edu

***Abstract*—This paper considers over-the-air federated learning (OTA-FL). OTA-FL exploits the superposition property of the wireless medium, and performs model aggregation over the air for free. Thus, it can greatly reduce the communication cost incurred in communicating model updates from the edge devices. In order to fully utilize this advantage while providing comparable learning performance to conventional federated learning that presumes model aggregation via noiseless channels, we consider the joint design of transmission scaling and the number of local iterations at each round, given the power constraint at each edge device. We first characterize the training error due to such channel noise in OTA-FL by establishing a fundamental lower bound for general functions with Lipschitz-continuous gradients. Then, by introducing an adaptive transceiver power scaling scheme, we propose an over-the-air federated learning algorithm with joint adaptive computation and power control (ACPC-OTA-FL). We provide the convergence analysis for ACPC-OTA-FL in training with non-convex objective functions and heterogeneous data. We show that the convergence rate of ACPC-OTA-FL matches that of FL with noise-free communications.**

## I. Introduction

In recent years, advances in machine learning (ML) have achieved astonishing successes in many applications that transform our society, e.g., in computer vision, natural language processing, and robotics. Traditionally, ML training tasks often reside in cloud-based large data-centers that process training data in a centralized fashion. However, due to the rapidly increasing demands for training data, high latency and costs of data transmissions, as well as data privacy/security concerns, aggregating all data to the cloud for ML training is unlikely to remain feasible. To address these challenges, federated learning (FL) [1] has recently emerged as a prevailing distributed ML paradigm. FL employs multiple clients, typically deployed over wireless edge networks, to locally train a learning model and exchange only intermediate updates between the server and clients. FL provides better avenues for privacy protection by avoiding the transmission of local data, while also being able to leverage parallel clients computation for training speedup.

However, FL also inherits many design challenges of distributed ML. One of the main challenges of FL stems from the communication constraint in the iterative FL learning process, particularly in resource (bandwidth and power)-limited wireless FL systems [2]–[4]. To receive the update information from multiple clients in each round, the conventional wisdom is to use orthogonal spectral or temporal channels for each client and avoid interference among the clients. However, this is neither desirable (since as the number of clients increases, the available rate per edge device decreases, lengthening the communication duration), nor necessary (since only the aggregated model updates is needed at the server) in FL.

Over-the-air FL (OTA-FL) has recently emerged as an effective approach in that it exploits the superposition property of the wireless medium to perform model aggregation "for free" by allowing simultaneous transmission of all clients' updates [5]–[7]. Specifically, under OTA-FL, the server directly recovers a noisy aggregation of the clients' model updates that transmit in the same spectral-temporal channel, rather than trying to decode each client's model update first in orthogonal spectral or temporal channels. As a result, OTA-FL dramatically reduces the communication costs and overheads from collecting the update from each client, and accordingly enjoys better *communication parallelism* regardless of the number of clients.

However, despite of the aforementioned salient features in terms of communication efficiency, several issues remain in OTA-FL. First, the convergence analysis of OTA-FL often assumes noise-free communications (see Section II for more in-depth discussions). Moreover, existing works on OTA-FL have not considered *data heterogeneity*, (i.e., datasets among clients are non-i.i.d. and with unbalanced sizes) and *system heterogeneity* (i.e., the computation and communication capacities varies among clients and could be time-varying [2], [3] simultaneously). Therefore, a fundamental question in OTA-FL system design is: *how to develop an efficient OTA-FL training algorithm that can handle both data and system heterogeneity under noisy channels.*

In this paper, we answer this question by studying the impact of channel noise on OTA-FL and proposing an over-the-air federated learning algorithm with joint adaptive computation and power control (ACPC-OTA-FL) for edge devices with heterogeneous capabilities. Our main contributions are summarized as follows:

- We first characterize the training error of the conventional OTA-based FedAvg algorithm [1] by establishing a lower bound of the convergence error under Gaussian multiple access channels (MAC) for general functions with Lipschitz continuous gradients. Our lower bound indicates that there is a non-vanishing convergence error due to channel noise compared with those convergence results of OTA-FL under

This work is supported in part by NSF-2112471 (AI-EDGE).

the noise-free assumption. This insight motivates us to propose our ACPC-OTA-FL algorithm that considers local computation and power control co-design to best utilize the power resources at the edge devices.

- Our ACPC-OTA-FL algorithm allows each client to (in a distributed manner) adaptively determine its **transmission power level and number of local update steps** to fully utilize the computation and communication resources. We show that, even with non-i.i.d. and unbalanced datasets, ACPC-OTA-FL converges to a stationary point with an $\mathcal{O}(1/\sqrt{mT})$ convergence rate for nonconvex objective functions in training, where $m$ is the number of clients and $T$ is the total number of training iterations. This result further implies a linear speedup in the number of clients and matches that of the noise-free FedAvg algorithm.

## II. Related Work

OTA-FL utilizes over-the-air computation through analog transmission in wireless MAC [5], [6], [8]–[13]. Despite the advantage of high scalability for large amount of clients, existing works on OTA-FL [9]–[13] have empirically shown that the channel noise substantially degrades the learning performance. Therefore, theoretically quantifying the impact of channel noise on convergence needs more in-depth investigation (See Section IV).

To mitigate the impacts of channel noise under limited transmission power constraints, one popular approach is to utilize uniform transceiver scaling for all clients. For example, references [11], [12], [14] have considered an identical sparsity pattern to reduce communication overhead. Reference [15] has proposed a new learning rate scheme that considers the quality of the gradient estimation; Reference [16] has developed a uniform-forcing transceiver scaling for OTA function computation, while the work in [17] has studied the optimal power control problem by taking gradient statistics into account. Reference [18] has proposed an uniform transceiver scaling by considering data heterogeneity. A common approach of these existing works above is to formulate the power control problem separately to satisfy the power constraints after the local computation at clients. Moreover, data and system heterogeneity and adapting the power resources for computation is not tied to transmission power control, despite the fact that each edge device is constrained in total power that is needed for both computation and communication. Due to the coupling of computation and communication processes, a joint adaptive computation and power control for OTA-FL is necessary in order to better mitigate the combined impacts of channel noise, power constraints, as well as data and system heterogeneity, which constitutes the main goal of this paper.

## III. System Model

In this section, we first introduce our OTA-FL model in Section III-A and then the communication model in Section III-B.

### A. Over-the-Air Federated Learning Model

We consider a FL system with one server and $m$ clients in total. Each client $i \in [m]$ contains a private local dataset $D_i$. Each local dataset $D_i$ is i.i.d. sampled from a distribution $\mathcal{X}_i$. In this paper, we consider non-i.i.d. datasests across users, i.e., $\mathcal{X}_i \neq \mathcal{X}_j$ if $i \neq j, \forall i, j \in [m]$. The goal of FL is to collaboratively train a global learning model by solving an optimization problem in the form of:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i\in[m]} \alpha_i F_i(\mathbf{x}, D_i), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the model parameter, $\alpha_i = \frac{|D_i|}{\sum_{i\in[m]}|D_i|}$ denotes the proportion of client $i$'s dataset size in the entire population. In this paper, we consider the unbalanced data setting: $\alpha_i \neq \alpha_j$ if $i \neq j$. In (1), $F_i(\mathbf{x}, D_i) \triangleq \frac{1}{|D_i|}\sum_{\xi_j^i \in D_i} F(\mathbf{x}, \xi_j^i)$ is the local objective function, where $\xi_j^i$ is the $j$-th sample in $D_i$. In FL, $F_i(\mathbf{x}, D_i)$ is assumed to be non-convex in general.

In each communication round $t$, the server broadcasts the latest global model parameter $\mathbf{x}_t$ to each client. Upon receiving $\mathbf{x}_t$, client $i$ performs local computations with $\mathbf{x}_{t,0}^i = \mathbf{x}_t$:

$$\mathbf{x}_{t,k+1}^i = \mathbf{x}_{t,k}^i - \eta \nabla F(\mathbf{x}_{t,k}^i, \xi_{t,k}^i), \quad k = 0, \ldots, \tau_t^i - 1, \tag{2}$$

where $\tau_t^i$ denotes the total number of local steps at client $i$ in round $t$ and $\xi_{t,k}^i$ is a random data sample used by client $i$ in step $k$ in round $t$. We note that one key feature of the OTA-FL model in this paper is that **we allow $\tau_t^i$ to be time-varying and device-dependent**. While this makes the OTA-FL model more practical and flexible, it also introduces an extra dimension of challenges in algorithmic design and convergence analysis.

After finishing the local iterations, each client returns the update of model parameters back to server. Upon the reception of returned updates from all clients, the server aggregates and updates the global model $\mathbf{x}_t$. In OTA-FL, the aggregation process on the server side happens automatically over-the-air thanks to the superposition property of the wireless medium. The specific communication model will be described in Section III-B.

### B. Communication Model

We consider an OTA-FL system in which the server broadcasts to all clients in a downlink channel and the clients transmit to the server through a common uplink channel synchronously. We assume an error-free downlink when broadcasting the global model. This is a reasonable assumption when the server has access to more power and bandwidth resources compared to edge devices. As a result, each client receives an error-free global model parameter $\mathbf{x}_t$ for its local computation in the beginning of each round $t$, i.e., $\mathbf{x}_{t,0}^i = \mathbf{x}_t$. For the uplink, we consider a Gaussian MAC, where the output the channel in each communication round $t$ is:

$$\mathbf{y}_t = \sum_{i\in[m]} h_t^i \mathbf{z}_t^i + \mathbf{w}_t. \tag{3}$$

In (3), $\mathbf{z}_t^i \in \mathbb{R}^d$ is the input from client $i$, $h_t^i$ is the channel gain of client $i$, and $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}_d)$ is an additive Gaussian channel noise.

We assume that each client $i$ has its channel state information (CSI) in each communication round $t$, i.e., $h_t^i$. This assumption is common in the OTA-FL literature since, in practice, CSI can be periodically estimated accurately between the server and each client. We also consider an instantaneous power constraint at each client in each communication round:

$$\|\mathbf{z}_t^i\|^2 \leq P_t^i, \quad \forall i \in [m], \forall t, \tag{4}$$

where $P_t^i$ is the maximum transmission power limit for client $i$ in communication round $t$.

## IV. Impacts of Channel Noise and System-Data Heterogeneity on OTA-FL

In Section IV-A, we first characterize the impact of the channel noise on OTA-FL when directly applying the standard FedAvg framework with SGD local updates without considering power control at each client. Then, in Section IV-B, we provide a concrete example to further illustrate the impact of channel noise coupled with heterogeneous numbers of local updates, i.e., system heterogeneity, on OTA-FL performance.

### A. Impact of Channel Noise on OTA-FL

To study the impact of channel noise, we first consider a general $L$-smooth objective function (i.e., having $L$-Lipschitz continous gradients) with a single local step, i.e., $\tau_t^i = 1, \forall i \in [m], t \in [T]$. Consequently, the channel output could be simplified as $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla F(\mathbf{x}_t, \xi_t) + \mathbf{w}_t$, where $\xi_t \triangleq \{\xi_t^i, \forall i \in [m]\}$ represents one collective data batch composed of random samples $\{\xi_t^i, \forall i\}$ from all clients. Then, we have the following theorem to characterize the impact of the channel noise on the OTA version of the FedAvg algorithm:

**Theorem 1** (Lower Bound for Gaussian Channel)**.** *Consider an OTA-FL system for training an $L$-smooth objective function $F(\mathbf{x})$ with an optimal solution $\mathbf{x}^*$. Supposed that each client uses local SGD updates that are subject to additive white Gaussian noise (AWGN), i.e., $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla F(\mathbf{x}_t, \xi_t) + \mathbf{w}_t$, where $\eta < \frac{1}{L}$ and $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}_d)$. Then, the sequence $\{\mathbf{x}_t\}$ satisfies the following recursive relationship:*

$$\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right] \geq \mathbb{E}\left[(1-\eta L)^2 \|\mathbf{x}_t - \mathbf{x}^*\|^2\right] + \eta^2 \sigma^2 + \sigma_c^2,$$

*which further implies the following lower bound for the training convergence:*

$$\lim_{t\to\infty} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|^2\right] \geq \frac{\eta^2\sigma^2 + \sigma_c^2}{1-(1-\eta L)^2}, \tag{5}$$

*where the stochastic gradient noise is assumed to be Gaussian, i.e., $\nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$.*

*Proof Sketch.* By assuming independent stochastic gradient noise and channel noise, we could decouple these noise terms and thus producing an iteration relation of $\|\mathbf{x}_t - \mathbf{x}_*\|$ by $L$-smoothness with proper learning rate $\eta < \frac{1}{L}$. As the channel noise exists in every round, such noise variance term $\sigma_c^2$ is non-vanishing even for infinitely many rounds. Due to space limitation, we refer readers to [19] for the complete proof. □

Theorem 1 suggests that there is a non-vanishing convergence error due to the Gaussian MAC noise when only standard SGD updates are used locally at each client. This motivates us to develop joint adaptive computation and power control for OTA-FL to mitigate the MAC noise effect.

### B. Impacts of System-Data Heterogeneity on OTA-FL

As discussed in above Section III-A, the FL optimization problem considered in this paper contains non-convex objective function, heterogeneous (non-i.i.d.) data, and different number of local updates $\tau_t^i$ at each client. As shown in previous work [20], different number of local steps (or optimization processes) among clients introduce objective inconsistency, rendering potentially arbitrary deviation from optimal solutions in conventional FL. Next, we show that similar negative impacts of data heterogeneity and different number of local steps also affect the OTA-FL performance even under proper power control. This insight further motivates the need for a *joint* adaptive computation and power control.

Here, we use the GD method for each client's local update and omit the channel noise for now to clearly characterize the above factors in FL. That is, each client $i \in [m]$ takes local steps as follows:

$$\mathbf{x}_{t,k+1}^i = \mathbf{x}_{t,k}^i - \eta \nabla F_i(\mathbf{x}_{t,k}^i), \quad \forall k \in [\tau_i]. \tag{6}$$

Then, a power control factor at client $i$ is applied as follows: $\mathbf{z}_t^i = \beta_i \left(\mathbf{x}_{t,\tau_i}^i - \mathbf{x}_{t,0}^i\right)$. The aggregation with power control factor $\beta$ on the server side can be written as

$$\mathbf{x}_{t+1} - \mathbf{x}_t = \sum_{i=1}^m \frac{\beta_i}{\beta} \left(\mathbf{x}_{t,\tau_i}^i - \mathbf{x}_{t,0}^i\right). \tag{7}$$

We now consider the following OTA-FL example:

**Example 1** (Deviation of Objective Value under Disjoint Power Control and System-Data Heterogeneity)**.** Consider quadratic objective functions:

$$F_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}_t \mathbf{H}_i \mathbf{x} - \mathbf{e}_i^T \mathbf{x} + \frac{1}{2}\mathbf{e}_i^T \mathbf{H}_i^{-1} \mathbf{e}_i, \forall i \in [m], \text{ and}$$

$$F(\mathbf{x}) = \sum_{i=1}^m \alpha_i F_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}_t \bar{\mathbf{H}} \mathbf{x} - \bar{\mathbf{e}_i}^T \mathbf{x} + \frac{1}{2}\sum_{i=1}^m \alpha_i \mathbf{e}_i^T \mathbf{H}_i^{-1} \mathbf{e}_i,$$

where $\mathbf{H}_i \in \mathbb{R}^{d\times d}$ is invertible, $\bar{\mathbf{H}} \triangleq \sum_{i=1}^m \alpha_i \mathbf{H}_i$, $\mathbf{e}_i \in \mathbb{R}^d$ is an arbitrary vector, and $\bar{\mathbf{e}} \triangleq \sum_{i=1}^m \alpha_i \mathbf{e}_i$. Let each client take $\tau_i$ GD steps. Then, the generated sequence $\{\mathbf{x}_t\}$ satisfies:

$$\lim_{t\to\infty} \mathbf{x}_t = \hat{\mathbf{x}}, \tag{8}$$

where the limit point $\hat{\mathbf{x}}$ can be computed as:

$$\hat{\mathbf{x}} = \left[\sum_{i=1}^m \frac{\beta_i}{\beta}\left[\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_i)^{\tau_i}\right] \mathbf{H}_i^{-1} \mathbf{H}_i\right]^{-1} \times \left[\sum_{i=1}^m \frac{\beta_i}{\beta}\left[\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_i)^{\tau_i}\right] \mathbf{H}_i^{-1} \mathbf{e}_i\right]. \tag{9}$$

For the quadratic objective function $F(\mathbf{x})$, the closed-form solution is $\mathbf{x}^* = \bar{\mathbf{H}}^{-1}\bar{\mathbf{e}}$. Comparing $\mathbf{x}^*$ and $\hat{\mathbf{x}}$, we can see a deviation that depends on problem hyper-parameter $\mathbf{H}$, learning rate $\eta$, local step numbers $\tau_i$, factor $\beta_i$ and $\beta$. □

**Algorithm 1** Adaptive Over-the-Air Federated Learning.

**Init: global mode** $\mathbf{x}_0$**.**
**for** $t = 0, \ldots, T-1$ **do**
  Server broadcasts latest global model $\mathbf{x}_t$ to each device.
  **for** each client $i \in [m]$ **do**
    Each client locally and adaptively trains model via SGD (10) under power constraints and transmits $\delta_t^i$ by transmission scaling as in (11).
  **end for**
  The server aggregates and updates global model by receiver rescaling (12).
**end for**

Due to space limitation, we provide the proof details of this example in our online technical report [19]. It can be seen from this example that the complex coupling between power control and system-data heterogeneity renders a highly non-trivial OTA-FL power control and algorithmic design to guarantee convergence to an optimal solution. This further motivates our OTA-FL algorithm design with *joint* adaptive computation and power control in Section V.

## V. Algorithm Design

To address the negative impacts of channel noise and system-data heterogeneity in Section IV, we propose an over-the-air federated learning algorithm with joint adaptive computation and power control (ACPC-OTA-FL) as shown in Algorithm 1. The basic idea of our ACPC-OTA-FL algorithm is to utilize a time-varying dynamic number of local SGD steps at each client under the instantaneous power constraint at this particular client. Specifically, given a server-side power control scaling factor $\beta_t$, each client $i$ performs local SGD steps as follows:

$$\mathbf{x}_{t,k+1}^i = \mathbf{x}_{t,k}^i - \eta \nabla F(\mathbf{x}_{t,k}^i, \xi_{t,k}^i), \quad k = 1, \ldots, \tau_i, \tag{10}$$

under the time-varying and client-dependent power constraints $P_t^i$ such that $\|\delta_t^i\|^2 \leq P_t^i$, where $\delta_t^i = \beta_t^i(\mathbf{x}_{t,\tau_t^i}^i - \mathbf{x}_{t,0}^i)$ and

$$\beta_t^i = \frac{\beta_t \alpha_i}{\tau_t^i}. \tag{11}$$

For simplicity, we ignore fading for now. The uplink channel output is $\sum_{i=1}^m \delta_t^i + \mathbf{w}_t$. With power control rescaling on the server side, the global model is aggregated and updated as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sum_{i=1}^m \frac{\delta_t^i}{\beta_t} + \tilde{\mathbf{w}}_t. \tag{12}$$

As mentioned earlier, we assume a Gaussian channel noise $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}_d)$ and thus $\tilde{\mathbf{w}}_t \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_c^2}{\beta_t^2}\mathbf{I}_d)$. The novelty of our algorithm lies in utilizing a time-varying dynamic local steps to fully exploit the computation and communication power resources.

There are two advantages in our algorithm compared to previous works. First, we jointly consider the computation-communication co-design due to their complex coupling relationship as shown in Section IV-B. As a result, more powerful clients with more computation capacities and transmission power will execute more local update steps and have a large fraction in the server-side aggregation. This adaptive and client-dependent design is different from previous works [11], [12], [14]–[18], which considered the communication problem separately after finishing local update computation and used an uniform power control scaling factor without considering the heterogeneity among the clients. Second, our ACPC-OTA-FL algorithm alleviates the straggler (i.e., slow client) problem by allowing different local step numbers across clients in each communication round. Before providing the theoretical convergence result, we first state our assumptions:

**Assumption 1.** *(L-Lipschitz Continuous Gradient) There exists a constant $L > 0$, such that $\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and $i \in [m]$.*

**Assumption 2.** *(Unbiased Local Stochastic Gradients and Their Bounded Variance) Let $\xi_i$ be a random local data sample at client $i$. The local stochastic gradient is unbiased and has a bounded variance, i.e., $\mathbb{E}[\nabla F_i(\mathbf{x}, \xi_i)] = \nabla F_i(\mathbf{x})$, $\forall i \in [m]$, and $\mathbb{E}[\|\nabla F_i(\mathbf{x}, \xi_i) - \nabla F_i(\mathbf{x})\|^2] \leq \sigma^2$, where the expectation is taken over the local data distribution $\mathcal{X}_i$.*

**Assumption 3.** *(Bounded Stochastic Gradient) There exist a constant $G \geq 0$, such that the norm of each local stochastic gradient is bounded: $\mathbb{E}[\|\nabla F_i(\mathbf{x}, \xi_i)\|^2] \leq G^2$, $\forall i \in [m]$.*

Assumptions 1–2 are common in the analysis of SGD-based algorithms. Assumption 3 is also widely used in OTA-FL with non-i.i.d. datasets (e.g., [12], [18], [21]). With these three assumptions, we have the following convergence result:

**Theorem 2** (Convergence Rate)**.** *Let $\{\mathbf{x}_t\}$ be the global model generated by Algorithm 1. Under Assumptions 1- 3 and a constant learning rate $\eta_t = \eta, \forall t \in [T]$, it holds that:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F(\mathbf{x}_t)\|^2 \leq \underbrace{\frac{2\left(F(\mathbf{x}_0) - F(\mathbf{x}_*)\right)}{T\eta}}_{\text{optimization error}} + \underbrace{L\eta\sigma^2 \sum_{i=1}^m \alpha_i^2}_{\text{statistical error}} + \underbrace{mL^2\eta^2 G^2 \sum_{i=1}^m (\alpha_i)^2 (\tau_i)^2}_{\text{local update error}} + \underbrace{\frac{L\sigma_c^2}{\eta\beta^2}}_{\text{channel noise error}},$$

*where $(\tau_i)^2 = \frac{\sum_{t=0}^{T-1}\left(\tau_t^i\right)^2}{T}$ and $\frac{1}{\beta^2} = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\beta_t^2}$.*

*Proof Sketch.* In each round, the average global model update is the same as noise-free FedAvg since the channel noise is independent Gaussian with zero mean. Thus, we can decouple the channel noise term as an extra error scaled by $\frac{\sigma_c^2}{\beta_t^2}$ when calculating the function descent $(F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t))$ in each round by the $L$-smoothness. Then, the technical challenge lies in heterogeneous local steps across clients. By simulating mini-batch SGD method, we could further bound the difference $\frac{1}{2}\eta_t \mathbb{E}_t[\|\sum_{i=1}^m \frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i))\|^2] \leq \frac{1}{2}\eta_t^3 mL^2 \sum_{i=1}^m (\alpha_i)^2 \left(\tau_t^i\right)^2 G^2$, which accounts for the size of dataset, data heterogeneity and different local steps. The above

two terms correspond to channel noise error and local update error, respectively. Following the classic analysis for SGD-based methods, the optimization error and statistical error could be similarly derived, and the final convergence result naturally follows. Due to space limitation, we relegate the full proof to our online technical report in [19]. □

Theorem 2 characterizes four sources of errors that affect the convergence rate: 1) the optimization error dependent on the distance between the initial guess and optimal objective value; 2) the statistical error due to the use of stochastic gradients rather than full gradients; 3) local update error from local update steps coupled with data heterogeneity; and 4) channel noise error from over-the-air transmissions. Among these four errors, only the optimization error (first term) vanishes as the total number of iterations $T$ gets large, while other three terms are independent of $T$.

Similar to classic SGD or FedAvg convergence analysis, diminishing learning rates $\mathcal{O}(\frac{1}{\sqrt{T}})$ can be used to remove the statistical and local update errors and obtain a convergence error bound $\min_{t\in[T]} \mathbb{E}\|\nabla F(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{\sqrt{T}})$. To mitigate the channel noise error, the parameter $\beta$ needs to be chosen judiciously. Given $\delta_t^i$ in communication round $t$, we can set $\beta_t^2 = \min_{i\in[m]}\{\frac{P_t^i(\tau_t^i)^2}{\|\delta_t^i\|^2\alpha_i^2}\}$. If the $\delta_t^i$-information is unavailable, we can choose $\|\delta_t^i\|^2 \leq \eta_t^2(\tau_t^i)^2G^2$ by its upper bound, and thus $\beta_t^2 = \frac{P_t}{\alpha_i^2\eta_t^2G^2}$, where $P_t = \min_{i\in[m]} P_t^i$. For the special case with $P = P_t^i, \forall i,t$, $\alpha_i = \frac{1}{m}$ (balanced datasets), and identical local steps $\tau_t^i = \tau, \forall i,t$, the channel noise error (the fourth term) becomes $\frac{\eta\sigma_c^2G^2}{Pm^2}$, and the following result immediately follows from Theorem 2:

**Corollary 1** (Convergence Rate)**.** *Let* $\alpha_i = \frac{1}{m}, \tau_i = \tau, \eta = \frac{\sqrt{m}}{\sqrt{T}}, \beta^2 = \frac{m}{\eta^2}$*, the convergence rate of ACPC-OTA-FL under the special case above is* $\mathcal{O}(\frac{\sigma^2+1}{\sqrt{mT}}) + \mathcal{O}(\frac{m\tau^2G^2}{T}) + \mathcal{O}(\frac{\sigma_c^2}{\sqrt{mT}})$.

Corollary 1 implies that, if $\tau \leq \frac{T^{1/4}}{m^{3/4}}$, a linear speedup in terms of the number of clients (i.e., $\mathcal{O}(\frac{1}{\sqrt{mT}})$) can be achieved, which shows the benefits of parallelism and matches the convergence rate of FedAvg in noise-free communication environment [22], [23].

Lastly, we note that it is straightforward to extend our results to fading channels with known CSI. Specifically, the adaptive computation and power control strategy for fading channels is to choose local steps $\tau_t^i$ such that $\|\delta_t^i\|^2 \leq P_t^i$, where $\delta_t^i = \beta_t^i\left(\mathbf{x}_{t,\tau_i}^i - \mathbf{x}_{t,0}^i\right), \beta_t^i = \frac{\beta_t\alpha_i}{\tau_t^i h_t^i}$, and $P_t^i > 0$ represents the maximum transmission power for client $i$ in round $t$. Under this joint computation and power control, the received signal remains the same as that in the non-fading OTA-FL setting. Thus, the same convergence results in Theorem 2 and Corollary 1 continue to hold.

## VI. Numerical Results

In this section, we conduct numerical experiments to verify our theoretical results using logistic regression on the MNIST dataset [24]. Following the same procedure as in existing works [1], [21], [23], we distribute the data evenly to $m = 10$ clients in a label-based partition to impose data heterogeneity across the clients, where the heterogeneity level can be characterized by a parameter $p$. As the MNIST dataset contains 10 classes of labels in total, $p = 10$ represents the i.i.d. case. The smaller the $p$-value, the more heterogeneous the data across clients. We simulate Gaussian MAC with signal-to-noise ratios (SNRs) of $-1$ dB, 10 dB and 20 dB.

TABLE I
Logistic regression test Accuracy (%) for ACPC-OTA-FL compared with COTAF and FedAvg on the MNIST dataset.

| Non-IID Level | Algorithm | Signal-to-Noise Ratio | | |
|---|---|---|---|---|
| | | -1 dB | 10 dB | 20 dB |
| $p=1$ | **ACPC-OTA-FL** | **78.22** | **89.08** | **89.54** |
| | COTAF | 46.55 | 65.54 | 85.92 |
| | FedAvg | 67.49 | 68.08 | 67.65 |
| $p=2$ | **ACPC-OTA-FL** | **81.89** | **89.58** | **90.43** |
| | COTAF | 63.59 | 78.80 | 86.57 |
| | FedAvg | 71.55 | 78.03 | 79.86 |
| $p=5$ | **ACPC-OTA-FL** | **86.48** | **90.64** | **91.20** |
| | COTAF | 79.64 | 86.52 | 90.82 |
| | FedAvg | 74.76 | 82.84 | 85.40 |
| $p=10$ | **ACPC-OTA-FL** | **86.21** | **90.75** | **91.08** |
| | COTAF | 86.43 | 91.08 | 92.63 |
| | FedAvg | 76.26 | 84.94 | 88.11 |

We illustrate the test accuracy of ACPC-OTA-FL compared with COTAF [18] and FedAvg [1] in Table I. Two key observations are in order: 1) Test accuracy drops significantly by directly applying FedAvg algorithm to wireless OTA-FL (up to 20% accuracy drop) under large channel noise, which validates our Theorem 1 and is consistent with existing works [9]–[11]; and 2) Under power control, both ACPC-OTA-FL and COTAF could alleviate the channel noise impacts in the i.i.d. data setting ($p = 10$). But in the highly heterogeneous data ($p = 1, 2$) and/or low SNR settings, our ACPC-OTA-FL algorithm outperforms COTAF by a large margin. For example, when SNR $= -1$ dB and $p = 1$, ACPC-OTA-FL improves the test accuracy by 31.76% and 10.73% compared to COTAF and FedAvg, respectively. The intuition is that the gradient returned from the clients vary dramatically in highly heterogeneous data settings, and thus utilizing an adaptive local steps under limited power constraints allows each client to fully exploit both computation and communication resources.

## VII. Conclusion

In this paper, we considered the joint adaptive local computation (number of local steps) and power control for OTA-FL. We first characterized the training error due to channel noise for conventional OTA-FL by establishing a fundamental lower bound for general objective functions with Lipschitz-continuous gradients. This motivated us to propose an over-the-air federated learning algorithm with joint adaptive computation and power control (ACPC-OTA-FL) to mitigate the impacts of channel noise on the learning performance, while taking the device heterogeneity into consideration. We analyzed the convergence of ACPC-OTA-FL with non-convex objective functions and heterogeneous data, and shown that the convergence rate of ACPC-OTA-FL matches that of FedAvg with noise-free communications.

## REFERENCES


[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[3] H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1, 2021.

[4] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[5] O. Abari, H. Rahul, and D. Katabi, "Over-the-air function computation in sensor networks," *arXiv preprint arXiv:1612.02307*, 2016.

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

[7] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive iot," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 57–65, 2021.

[8] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning based on over-the-air computation," in *ICC 2019-2019 IEEE international conference on communications (ICC)*. IEEE, 2019, pp. 1–6.

[9] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2019.

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

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

[12] ——, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[13] ——, "Over-the-air machine learning at the wireless edge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[14] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2120–2135, 2020.

[15] H. Guo, A. Liu, and V. K. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 197–210, 2020.

[16] L. Chen, X. Qin, and G. Wei, "A uniform-forcing transceiver design for over-the-air function computation," *IEEE Wireless Communications Letters*, vol. 7, no. 6, pp. 942–945, 2018.

[17] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–6.

[18] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, 2021.

[19] Y. Haibo, Q. Peiwen, L. Jia, and Y. Aylin, "Over-the-air federated learning with jointadaptive computation and power control," The Ohio State University, Tech. Rep., February 2022. [Online]. Available: https://kevinliu-osu.github.io/publications/OTA-FL-TR.pdf

[20] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[21] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJxNAnVtDS

[22] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5693–5700.

[23] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-IID federated learning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=jDdzh5ul-d

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

## VIII. Proof

**Theorem 1** (Lower Bound for Gaussian Channel)**.** *Consider an OTA-FL system for training an $L$-smooth objective function $F(\mathbf{x})$ with an optimal solution $\mathbf{x}^*$. Supposed that each client uses local SGD updates that are subject to additive white Gaussian noise (AWGN), i.e., $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla F(\mathbf{x}_t, \xi_t) + \mathbf{w}_t$, where $\eta < \frac{1}{L}$ and $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_c^2\mathbf{I}_d)$. Then, the sequence $\{\mathbf{x}_t\}$ satisfies the following recursive relationship:*

$$\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right] \geq \mathbb{E}\left[(1-\eta L)^2 \|\mathbf{x}_t - \mathbf{x}^*\|^2\right] + \eta^2\sigma^2 + \sigma_c^2,$$

*which further implies the following lower bound for the training convergence:*

$$\lim_{t\to\infty} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|^2\right] \geq \frac{\eta^2\sigma^2 + \sigma_c^2}{1-(1-\eta L)^2}, \tag{5}$$

*where the stochastic gradient noise is assumed to be Gaussian, i.e., $\nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$.*

*Proof.*

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right] &= \mathbb{E}\left[\|\mathbf{x}_t - \eta\nabla F(\mathbf{x}_t, \xi_t) + \mathbf{w}_t - \mathbf{x}^*\|^2\right] && (13)\\
&= \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^* - \eta\nabla F(\mathbf{x}_t)\|^2\right] + \mathbb{E}\left[\|\eta\nabla F(\mathbf{x}_t, \xi_t) - \eta\nabla F(\mathbf{x}_t)\|^2\right] + \mathbb{E}\left[\|\mathbf{w}_t\|^2\right] && (14)\\
&= \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^* - \eta\left(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*)\right)\|^2\right] + \eta^2\sigma^2 + \sigma_c^2 && (15)\\
&\geq \mathbb{E}\left[\left(\|\mathbf{x}_t - \mathbf{x}^*\| - \eta\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*)\|\right)^2\right] + \eta^2\sigma^2 + \sigma_c^2 && (16)\\
&\geq \mathbb{E}\left[(1-\eta L)^2\|\mathbf{x}_t - \mathbf{x}^*\|^2\right] + \eta^2\sigma^2 + \sigma_c^2 && (17)
\end{aligned}$$

□

*

*Proof.* Consider quadratic objective functions:

$$F_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}_t\mathbf{H}_i\mathbf{x} - \mathbf{e}_i^T\mathbf{x} + \frac{1}{2}\mathbf{e}_i^T\mathbf{H}_i^{-1}\mathbf{e}_i, \tag{18}$$

$$F(\mathbf{x}) = \sum_{i=1}^m \alpha_i F_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}_t\bar{\mathbf{H}}\mathbf{x} - \bar{\mathbf{e}_i}^T\mathbf{x} + \frac{1}{2}\sum_{i=1}^m \alpha_i\mathbf{e}_i^t\mathbf{H}_i^{-1}\mathbf{e}_i, \tag{19}$$

where $\mathbf{H}_i \in \mathbb{R}^{d\times d}$ is invertible matrix, $\bar{\mathbf{H}} = \sum_{i=1}^m \alpha_i\mathbf{H}_i$, $\mathbf{e}_i \in \mathbb{R}^d$ is arbitrary vector, and $\bar{\mathbf{e}} = \sum_{i=1}^m \alpha_i\mathbf{e}_i$. It is easy to show the optimum for each local and global objective function are $\mathbf{x}_i^* = \mathbf{H}_i^{-1}\mathbf{e}_i, \forall i \in [m]$ and $\mathbf{x}^* = \bar{\mathbf{H}}^{-1}\bar{\mathbf{e}}$.

The local update for each client $i \in [m]$ is as follows:

$$\begin{aligned}
\mathbf{x}_{t,k+1}^i &= \mathbf{x}_{t,k}^i - \eta\left[\mathbf{H}_i\mathbf{x}_{t,k}^i - \mathbf{e}_i\right] && (20)\\
&= \left(\mathbf{I} - \eta\mathbf{H}_i\right)\mathbf{x}_{t,k}^i + \eta\mathbf{e}_i && (21)
\end{aligned}$$

Then we have the recursive equation for one local step by rearranging 21:

$$\mathbf{x}_{t,k+1}^i - \mathbf{c}_t^i = \left(\mathbf{I} - \eta\mathbf{H}_i\right)\left(\mathbf{x}_{t,k}^i - \mathbf{c}_t^i\right), \tag{22}$$

where $\mathbf{c}_t^i = \mathbf{H}_i^{-1}\mathbf{e}_i$.

$$\begin{aligned}
\mathbf{x}_{t,\tau_i}^i - \mathbf{c}_t^i &= \left(\mathbf{I} - \eta\mathbf{H}_i\right)^{\tau_i}\left(\mathbf{x}_{t,0}^i - \mathbf{c}_t^i\right), && (23)\\
\mathbf{x}_{t,\tau_i}^i - \mathbf{x}_{t,0}^i &= \left[\left(\mathbf{I} - \eta\mathbf{H}_i\right)^{\tau_i} - \mathbf{I}\right]\mathbf{H}_i^{-1}\left(\mathbf{e}_i - \mathbf{H}_i\mathbf{x}_{t,0}^i\right) && (24)\\
&= \mathbf{K}_i(\eta)\left(\mathbf{e}_i - \mathbf{H}_i\mathbf{x}_{t,0}^i\right), && (25)
\end{aligned}$$

where we define $\mathbf{K}_i(\eta) = \left[\mathbf{I} - \left(\mathbf{I} - \eta\mathbf{H}_i\right)^{\tau_i}\right]\mathbf{H}_i^{-1}$.

Aggregation:

$$\begin{aligned}
\mathbf{x}_{t+1} - \mathbf{x}_t &= \sum_{i=1}^m \frac{\beta_i}{\beta}\left(\mathbf{x}_{t,\tau_i}^i - \mathbf{x}_{t,0}^i\right) && (26)\\
&= \sum_{i=1}^m \frac{\beta_i}{\beta}\mathbf{K}_i(\eta)\left(\mathbf{e}_i - \mathbf{H}_i\mathbf{x}_{t,0}^i\right), && (27)
\end{aligned}$$

$$\mathbf{x}_{t+1} = \left[\mathbf{I} - \sum_{i=1}^{m} \frac{\beta_i}{\beta} \mathbf{K}_i(\eta)\mathbf{H}_i\right] \mathbf{x}_t + \sum_{i=1}^{m} \frac{\beta_i}{\beta} \mathbf{K}_i(\eta)\mathbf{e}_i, \tag{28}$$

$$\mathbf{x}_{t+1} - \hat{\mathbf{x}} = \left[\mathbf{I} - \sum_{i=1}^{m} \frac{\beta_i}{\beta} \mathbf{K}_i(\eta)\mathbf{H}_i\right] (\mathbf{x}_t - \hat{\mathbf{x}}), \tag{29}$$

$$\mathbf{x}_t = \left[\mathbf{I} - \sum_{i=1}^{m} \frac{\beta_i}{\beta} \mathbf{K}_i(\eta)\mathbf{H}_i\right]^t (\mathbf{x}_0 - \hat{\mathbf{x}}) + \hat{\mathbf{x}}, \tag{30}$$

where $\hat{\mathbf{x}} := \left[\sum_{i=1}^m \frac{\beta_i}{\beta}\mathbf{K}_i(\eta)\mathbf{H}_i\right]^{-1}\left[\sum_{i=1}^m \frac{\beta_i}{\beta}\mathbf{K}_i(\eta)\mathbf{e}_i\right] = \left[\sum_{i=1}^m \frac{\beta_i}{\beta}\left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_i)^{\tau_i}\right]\mathbf{H}_i^{-1}\mathbf{H}_i\right]^{-1}\left[\sum_{i=1}^m \frac{\beta_i}{\beta}\left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_i)^{\tau_i}\right]\mathbf{H}_i^{-1}\mathbf{e}_i\right]$.

As $t$ goes to sufficiently large, we have

$$\lim_{t\to\infty} \mathbf{x}_t = \hat{\mathbf{x}}.$$

□

**Theorem 2** (Convergence Rate)**.** *Let $\{\mathbf{x}_t\}$ be the global model generated by Algorithm 1. Under Assumptions 1- 3 and a constant learning rate $\eta_t = \eta, \forall t \in [T]$, it holds that:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F(\mathbf{x}_t)\|^2 \leq \underbrace{\frac{2\left(F(\mathbf{x}_0) - F(\mathbf{x}_*)\right)}{T\eta}}_{\text{optimization error}} + \underbrace{L\eta\sigma^2 \sum_{i=1}^{m} \alpha_i^2}_{\text{statistical error}}$$

$$+ \underbrace{mL^2\eta^2G^2 \sum_{i=1}^{m} (\alpha_i)^2 (\tau_i)^2}_{\text{local update error}} + \underbrace{\frac{L\sigma_c^2}{\eta\beta^2}}_{\substack{\text{channel noise} \\ \text{error}}},$$

*where* $(\tau_i)^2 = \frac{\sum_{t=0}^{T-1}(\tau_t^i)^2}{T}$ *and* $\frac{1}{\beta^2} = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\beta_t^2}$.

*Proof.*

$$\mathbf{x}_{t+1} - \mathbf{x}_t = \sum_{i=1}^{m} \frac{\beta_i}{\beta}\left(\mathbf{x}_{t,\tau_t^i}^i - \mathbf{x}_{t,0}^i\right) + \tilde{\mathbf{w}}_t \tag{31}$$

$$= \sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\left(\mathbf{x}_{t,\tau_t^i}^i - \mathbf{x}_{t,0}^i\right) + \tilde{\mathbf{w}}_t \tag{32}$$

$$= \sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\eta_t \sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right) + \tilde{\mathbf{w}}_t \tag{33}$$

Due to L-smoothness, we have one step descent in expectation conditioned on $\mathbf{x}_t$,

$$\mathbb{E}_t[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq \langle \nabla F(\mathbf{x}_t), \mathbb{E}_t\left[\mathbf{x}_{t+1} - \mathbf{x}_t\right]\rangle + \frac{L}{2}\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] \tag{34}$$

$$= \left\langle \nabla F(\mathbf{x}_t), \sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\eta_t \sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\rangle + \frac{L}{2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{m} \frac{\alpha_i\eta_t}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right) + \tilde{\mathbf{w}}_t\right\|^2\right] \tag{35}$$

$$= -\eta_t\|\nabla F(\mathbf{x}_t)\|^2 + \eta_t\left\langle \nabla F(\mathbf{x}_t), \nabla F(\mathbf{x}_t) - \sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\rangle \tag{36}$$

$$+ \frac{L}{2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{m} \frac{\alpha_i\eta_t}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right) + \tilde{\mathbf{w}}_t\right\|^2\right] \tag{37}$$

$$= -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 - \frac{1}{2}\eta_t\left\|\sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 + \frac{1}{2}\eta_t\mathbb{E}_t\left\|\nabla F(\mathbf{x}_t) - \sum_{i=1}^{m} \frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{38}$$

$$+ \frac{L}{2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{m} \frac{\alpha_i\eta_t}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right) + \tilde{\mathbf{w}}_t\right\|^2\right] \tag{39}$$

$$= -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 - \frac{1}{2}\eta_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 + \frac{1}{2}\eta_t\mathbb{E}_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{40}$$

$$+ \frac{L\eta_t^2}{2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right)\right\|^2\right] + \frac{L\sigma_c^2}{2\beta_t^2} \tag{41}$$

$$\leq -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\eta_t\mathbb{E}_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{42}$$

$$+ \frac{L\eta_t^2}{2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i)\right) - \sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2\right] + \frac{L\sigma_c^2}{2\beta_t^2} \tag{43}$$

$$\leq -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\eta_t\mathbb{E}_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{44}$$

$$+ \frac{L\eta_t^2}{2}\sum_{i=1}^{m}\mathbb{E}_t\left[\left\|\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_{t,k}^i, \xi_{t,k}^i) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2\right] + \frac{L\sigma_c^2}{2\beta_t^2} \tag{45}$$

$$\leq -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\eta_t\mathbb{E}_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 + \frac{L\eta_t^2}{2}\sum_{i=1}^{m}\alpha_i^2\sigma^2 + \frac{L\sigma_c^2}{2\beta_t^2} \tag{46}$$

The first inequality holds if $\eta_t \leq \frac{1}{L}$.

$$\frac{1}{2}\eta_t\mathbb{E}_t\left\|\sum_{i=1}^{m}\frac{\alpha_i}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \leq \frac{1}{2}\eta_t m\sum_{i=1}^{m}\frac{(\alpha_i)^2}{(\tau_t^i)^2}\mathbb{E}_t\left\|\sum_{k=0}^{\tau_t^i-1}\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{47}$$

$$\leq \frac{1}{2}\eta_t m\sum_{i=1}^{m}\frac{(\alpha_i)^2}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\mathbb{E}_t\left\|\left(\nabla F_i(\mathbf{x}_t) - \nabla F_i(\mathbf{x}_{t,k}^i)\right)\right\|^2 \tag{48}$$

$$\leq \frac{1}{2}\eta_t mL^2\sum_{i=1}^{m}\frac{(\alpha_i)^2}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\mathbb{E}_t\left\|\left(\mathbf{x}_t - \mathbf{x}_{t,k}^i\right)\right\|^2 \tag{49}$$

$$\leq \frac{1}{2}\eta_t^3 mL^2\sum_{i=1}^{m}\frac{(\alpha_i)^2}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}\mathbb{E}_t\left\|\sum_{j=0}^{k}\nabla F_i(\mathbf{x}_{t,j}^i, \xi_{t,j}^i)\right\|^2 \tag{50}$$

$$\leq \frac{1}{2}\eta_t^3 mL^2\sum_{i=1}^{m}\frac{(\alpha_i)^2}{\tau_t^i}\sum_{k=0}^{\tau_t^i-1}k^2G^2 \tag{51}$$

$$\leq \frac{1}{2}\eta_t^3 mL^2\sum_{i=1}^{m}(\alpha_i)^2\left(\tau_t^i\right)^2G^2 \tag{52}$$

Plugging inequality (52) into (46), we have

$$\mathbb{E}_t[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq \langle\nabla F(\mathbf{x}_t), \mathbb{E}_t\left[\mathbf{x}_{t+1} - \mathbf{x}_t\right]\rangle + \frac{L}{2}\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] \tag{53}$$

$$\leq -\frac{1}{2}\eta_t\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\eta_t^3 mL^2\sum_{i=1}^{m}(\alpha_i)^2\left(\tau_t^i\right)^2G^2 + \frac{L\eta_t^2}{2}\sum_{i=1}^{m}\alpha_i^2\sigma^2 + \frac{L\sigma_c^2}{2\beta_t^2} \tag{54}$$

Rearranging and telescoping:

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t\mathbb{E}_t\|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{2\left(F(\mathbf{x}_0) - F(\mathbf{x}_T)\right)}{T} + mL^2\frac{1}{T}\sum_{t=0}^{T-1}\eta_t^3\sum_{i=1}^{m}(\alpha_i)^2\left(\tau_t^i\right)^2G^2 + L\sigma^2\sum_{i=1}^{m}\alpha_i^2\frac{1}{T}\sum_{t=0}^{T-1}\eta_t^2 + L\sigma_c^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\beta_t^2} \tag{55}$$

Let $\eta_t = \eta$ be constant learning rate, $(\tau_i)^2 = \frac{\sum_{t=0}^{T-1} (\tau_t^i)^2}{T}$ then we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{2\left(F(\mathbf{x}_0) - F(\mathbf{x}_T)\right)}{T\eta} + mL^2\eta^2G^2\sum_{i=1}^{m}(\alpha_i)^2\left(\tau_i\right)^2 + L\eta\sigma^2\sum_{i=1}^{m}\alpha_i^2 + \frac{L\sigma_c^2}{\eta\beta^2}, \tag{56}$$

where $\frac{1}{\beta^2} = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\beta_t^2}$. □