

Towards Federated Learning Over the Air: Why Scaling Up Helps?

Jiaqi Zhu^{*†}, Bikramjit Das[‡], Nikolaos Pappas[§], and Howard H. Yang^{*†}

^{*}Zhejiang University/University of Illinois Institute, Zhejiang University, Haining, China

[†]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

[‡]Engineering Systems and Design, Singapore University of Technology and Design, Singapore

[§]Dept. of Computer and Information Science, Linköping University, Linköping, Sweden

Abstract—Federated learning enables multiple clients to collaboratively train a common model while concurrently preserving data privacy. However, its performance is often constrained by limited communication resources, especially when the system encounters a large number of clients. Under those circumstances, integrating over-the-air computations into the model training procedure is considered an effective approach to coping with the communication bottleneck. Specifically, by uploading each client’s intermediate parameters via analog transmissions instead of digital ones, the system can dramatically extend the number of clients it simultaneously supports in each communication round. However, that is achieved at the expense of introducing channel distortions, particularly fading and noise, in the aggregated global parameter. To demystify these effects, the present paper develops a theoretical framework to analyze the performance of the over-the-air federated model training process. Our analysis unveils a three-fold benefit from system scaling up, i.e., as the number of participating clients increases: (i) the privacy leakage, quantified by the mutual information between each client’s locally possessed gradient and the edge’s globally aggregated one, substantially decreases, (ii) the impairment of small-scale fading disappears due to the channel hardening effect, and (iii) the convergence rate is enhanced as thermal noise and gradient estimation error can be reduced. To that end, it establishes over-the-air model training as a viable approach for implementing federated learning in scenarios with a large number of clients. We corroborate the theoretical findings with extensive experiments.

Index Terms—Federated learning, over-the-air computing, privacy leakage, convergence rate.

I. INTRODUCTION

Federated learning (FL) is an emerging distributed machine learning paradigm that enables collaborative training of a global model across multiple clients without exposing their private local data to a central server [1]. Although FL enhances data privacy by keeping local data in-device, the frequent exchange of parameters between the edge server and clients incurs significant communication overhead. This becomes a paramount performance bottleneck for FL, especially when the number of participating clients becomes large [2].

The work of J. Zhu and H. H. Yang was supported in part by the National Key R&D Program of China under Grant 2024YFE0200700, in part by the open research fund of National Mobile Communications Research Laboratory, and in part by the National Natural Science Foundation of China under Grant 62201504. The work of N. Pappas has been supported by the Swedish Research Council (VR), ELLIIT, and the European Union (ETHER, 101096526). (*Corresponding Author: Howard H. Yang*)

In response, a line of recent studies [3]–[5] suggested integrating over-the-air computations into the FL model training procedure, leveraging the superposition properties of radio waveforms for fast and scalable parameter aggregation. Unlike digital communication-based parameter uploading, analog transmission avoids the linear increase in spectrum resource consumption as the number of participating entities grows. Moreover, an increase in the number of clients can improve energy efficiency, thereby reducing the transmit power [5]; in addition, analog over-the-air computing bypasses the encoding (resp. decoding) and modulation (resp. demodulation) processes, significantly reducing the access latency [3], [4]. As a result, adopting over-the-air computations enables the implementation of large-scale edge learning systems with low-cost communication modules [6], [7].

However, these are achieved at the price of introducing additional distortions into the received signal. Inevitably corrupted by channel fading and thermal noise, the noisy gradient results in unstable training performance or even impedes the efficacy of model convergence. To that end, several works [8]–[10] proposed estimating the instantaneous channel gains of every client before each global transmission, such that the clients can adequately assert power control to compensate for the channel fading (and thermal noise). Unfortunately, when a large number of clients are present in the system (which is the appropriate situation for adopting over-the-air computations in FL training), accurate estimation of instant channel gains becomes exceptionally costly. This naturally leads to a crucial question: *Is instant channel estimation and the subsequent power control necessary for federated model training over the air in large-scale networks?* The answer from this paper is no.

While detrimental to training efficiency, channel distortions contribute to enhancing end-user privacy. Although privacy protection is emphasized as a salient feature of FL, studies have shown that it is not sufficient to ensure privacy [11], [12]. For instance, private data can still be inferred from individual updates through model inversion attack [12] or membership inference attack [13]. In contrast, over-the-air aggregation ensures potential eavesdroppers can only access the aggregated updates, thereby protecting the privacy of the participating clients. Existing research has identified this aspect [14]–[16], in which [14] proposed harnessing inherent receiver noise to ensure differential privacy (DP) against inference attacks,

with the number of clients being a key factor for maintaining a higher privacy level and [15] showed that channel noise offers “free” privacy when the privacy constraint level is below a certain threshold. Nevertheless, how much privacy is guaranteed or how much information the aggregated model updates leak about a single client’s local dataset remains unclear. The present paper develops an information-theoretic metric to measure privacy leakage in the context of over-the-air model aggregation.

In brief, although adopting over-the-air computations in model aggregation is originally regarded as an alternative to circumvent the communication bottleneck when the system needs to support a large number of clients, the key finding in this paper reveals that the reverse is also true, i.e., *the presence of a massive amount of clients enhances the performance of federated learning over the air*.

The technical contributions are summarized below.

- We quantify the privacy leakage of analog over-the-air model aggregation by deriving an analytical expression for the mutual information (MI) between each client’s locally possessed gradient and the globally aggregated one. The analysis reveals a two-fold benefit of model training over the air: On the one hand, distortions induced by analog transmissions reduce privacy leakage; on the other hand, with a large number of clients present, each participant’s MI approaches zero.
- We establish a concentration inequality for noisy gradient obtained from over-the-air model aggregation, which unveils a channel-hardening effect in the global averaging stage. Subsequently, we derive analytical expressions for the convergence rate for non-convex loss functions. The result shows that as the number of clients increases, the impairments from communication and estimation noise vanish, thus expediting a stable convergence process. In addition, the convergence can be further improved by appropriately choosing local batch size on the client side.
- We conduct extensive simulations to verify the theoretical findings. The experiments confirm that increasing the number of clients improves privacy protection and training efficiency, validating the benefits of scaling up the system.

Notations: Throughout the paper, bold lowercase letters represent column vectors. For a given vector \mathbf{w} , \mathbf{w}^\top denotes its transpose and $\|\mathbf{w}\|$ denotes its L-2 norm. The identity matrix of dimension $d \times d$ is denoted by \mathbf{I}_d . For any positive integer i , $[i]$ denotes the set of integers $\{1, 2, \dots, i\}$, and $\mathbf{w}_{(i)}$ denotes the i -th entry of the vector \mathbf{w} .

II. SYSTEM MODEL

A. Setting

We consider the federated edge learning system depicted in Fig. 1, which comprises an edge server and N clients, where $N \gg 1$. Every client n holds a local dataset $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ represent the data sample and the corresponding response, respectively. We

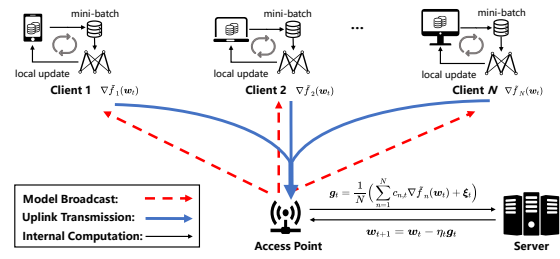


Fig. 1: An overview of the edge learning system.

assume the local datasets are statistically independent from each other. For simplicity, we assume the clients have equal-sized datasets, i.e., $m_n = M$, $\forall n \in [N]$.

The goal of all the entities in this system is to collaboratively train a statistical model over all the data samples from the clients without sharing their private data. More precisely, they need to jointly minimize the following objective function

$$f(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N f_n(\mathbf{w}) \quad (1)$$

where $f_n(\mathbf{w})$ is the local empirical risk of client n , given by

$$f_n(\mathbf{w}) = \frac{1}{m_n} \sum_{i=1}^{m_n} \ell(\mathbf{w}; \mathbf{x}_i, y_i) = \frac{1}{M} \sum_{i \in \mathcal{D}_n} \ell(\mathbf{w}; i) \quad (2)$$

in which $\ell(\mathbf{w}; \mathbf{x}_i, y_i)$ quantifies the loss associated with the sample pair (\mathbf{x}_i, y_i) , and we simplify its notation by $\ell(\mathbf{w}; i)$ for convenience. The optimal solution of (1) is commonly known as the empirical risk minimizer, denoted by

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}). \quad (3)$$

B. Analog Over-the-Air Model Training

The general procedure of federated model training under analog over-the-air computing has been detailed in [6]. We briefly describe it in this part for completeness. At the t -th round of global communication, the edge server broadcasts the current global model \mathbf{w}_t to all the clients. Subsequently, each client n initializes its local model as $\mathbf{w}_{n,t}^{(0)} = \mathbf{w}_t$, randomly shuffles its local dataset \mathcal{D}_n and splits it into $\lfloor \frac{M}{B} \rfloor$ mini batches, each containing B data samples. Then, the client performs E stochastic gradient descent steps based on the mini batches. More precisely, the local model of client n at the k -th local iteration, $k \in [E]$, is updated as

$$\mathbf{w}_{n,t}^{(k+1)} = \mathbf{w}_{n,t}^{(k)} - \eta_l \frac{1}{B} \sum_{j=1}^B \nabla \ell(\mathbf{w}_{n,t}^{(k)}; \theta_{n,t}^{(k)}(j)) \quad (4)$$

where η_l denotes the local learning rate and $\theta_{n,t}$ represents a permutation to the local dataset (of client n at the t -th global iteration). Upon the completion of local training, client n has its accumulated local gradient as follows:

$$\nabla \tilde{f}_n(\mathbf{w}_t) = \sum_{k=0}^{E-1} \frac{1}{B} \sum_{j=1}^B \nabla \ell(\mathbf{w}_{n,t}^{(k)}; \theta_{n,t}^{(k)}(j)). \quad (5)$$

After that, every client modulates its local gradient onto the magnitude of a set of common orthogonal baseband waveforms, in an entry-wise manner and simultaneously transmits the analog signal to the edge server. We consider the clients use power control to compensate for large-scale path loss, estimated via long-term averages of the received signal strength [17], while instantaneous fading remains unknown. Each client has a maximum transmit power per communication round.

Due to the superposition property of radio waveforms, the edge server can pass the received signal to matched filters and output the automatically aggregated (but distorted) gradient

$$\mathbf{g}_t = \frac{1}{N} \left(\sum_{n=1}^N c_{n,t} \nabla \tilde{f}_n(\mathbf{w}_t) + \boldsymbol{\xi}_t \right) \quad (6)$$

where $c_{n,t}$ denotes the channel fading experienced by client n and $\boldsymbol{\xi}_t \in \mathbb{R}^d$ represents the thermal noise vector. In this work, we assume the channel fading varies independently and identically distributed (i.i.d.) across the clients and communication rounds, with mean μ_c and variance σ_c^2 . Moreover, we model the thermal noise as an additive white Gaussian noise (AWGN) with variance σ_z^2 , i.e., $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \sigma_z^2 \mathbf{I}_d)$.

Using (6), the edge server updates the global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t \quad (7)$$

where η_t is the global learning rate. The global parameter will be broadcast to all the clients for the next round of local training. Such iterations repeat until the model converges.

Remark 1: Although each client modulates its raw gradient information directly onto the common waveform bases, the analog signals have to undergo a radio spectrum that distorts the original signals (through the effects of channel fading and thermal noise) before accumulating at the edge server. The edge server can only extract an automatically aggregated gradient from the received signal without accessing each client's information. In that respect, analog over-the-air computation constitutes a form of secure aggregation [18].

C. Performance Metric

We assess system performance from the perspective of privacy leakage and training efficiency. Specifically, we apply mutual information (MI), a metric that quantifies the shared information between two entities, to assess privacy leakage [19]. For a client n , the MI between its local gradient information and the global one in communication round t is defined as

$$I_N^{(t)} = \max_{n \in [N]} I \left(\nabla \tilde{f}_n(\mathbf{w}_t); \mathbf{g}_t \mid \{\mathbf{g}_p\}_{p \in [t-1]} \right). \quad (8)$$

We measure the training efficiency by the rate at which the (time-average) global gradient approaches zero with communication rounds (a.k.a. the convergence rate). Formally, after T rounds of global iterations, the convergence rate can be quantified by the following

$$R_N^{(T)} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\mathbf{w}_t)\|^2]. \quad (9)$$

III. ANALYSIS

A. Mutual Information

To focus on the analysis of privacy leakage, we adopt the honest-but-curious model [18], wherein the server honestly performs the required operations but may attempt to infer sensitive information about individual clients from the aggregated gradients. Consequently, we exclude scenarios involving malicious attackers or adversarial behavior.

We focus exclusively on the case where the local dataset \mathcal{D}_n are sampled i.i.d. from a common distribution to simplify the factors under consideration. This implies that conditioned on the received global model \mathbf{w}_t , the distribution of the local stochastic gradients $\nabla \tilde{f}_n(\mathbf{w}_t)$, $\forall n \in [N]$, are also i.i.d.. The following definition will be employed to quantify the MI.

Definition 1 (Independent Under Whitening): A random vector \mathbf{v} with mean $\boldsymbol{\mu}_v$ and a non-singular covariance matrix \mathbf{K}_v is said to be independent under whitening if the components of the whitened vector, $\hat{\mathbf{v}} = \mathbf{K}_v^{-1/2}(\mathbf{v} - \boldsymbol{\mu}_v)$, are independent random variables.

We are now ready to present the information leakage in a single communication round, which is detailed below.

Theorem 1: Let \mathcal{S}_g denote the set of subvectors of dimension d^* of $\nabla \tilde{f}_n(\mathbf{w}_t)$ that have a non-singular covariance matrix, where $d^* \leq d$. If there exists $\bar{\mathbf{g}} \in \mathcal{S}_g$ with mean $\boldsymbol{\mu}_t$ and non-singular covariance matrix \mathbf{K}_t that is independent under whitening, and $\mathbb{E}[\|\bar{\mathbf{g}}_{(i)}\|^4] < \infty$ for all $i \in [d^*]$, then $\exists C_{\bar{\mathbf{g}}} > 0$, with which we can upper bound $I_N^{(t)}$ as follows:

$$I_N^{(t)} \leq \frac{C_{\bar{\mathbf{g}}} d^*}{N-1} + \frac{1}{2} \sum_{i=1}^{d^*} \log \left(\frac{N \lambda_{t,i} + \sigma_z^2}{(N-1) \lambda_{t,i} + \sigma_z^2} \right) \quad (10)$$

where $\{\lambda_{t,i}\}_{i=1}^{d^*}$ represent the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_t = (\mu_c^2 + \sigma_c^2) \mathbf{K}_t + \sigma_c^2 \boldsymbol{\mu}_t \boldsymbol{\mu}_t^\top$.

Proof: Please refer to Appendix A. \square

An immediate result is the MI under noiseless transmission (in this case, $\mu_c = 1$ and $\sigma_c^2 = \sigma_z^2 = 0$), where the MI in the t -th communication round can be bounded by [19]

$$I_N^{(t)} \leq \frac{C_{\bar{\mathbf{g}}} d^*}{N-1} + \frac{d^*}{2} \log \left(\frac{N}{N-1} \right). \quad (11)$$

Remark 2: Since $\frac{N \lambda_{t,i} + \sigma_z^2}{(N-1) \lambda_{t,i} + \sigma_z^2} < \frac{N}{N-1}$, the result in the right-hand-side of (10) is smaller than that of (11), revealing that channel distortions introduced by over-the-air computations can enhance privacy protection.

B. Convergence Rate

In this part, we derive the convergence rate of the considered edge learning system. To facilitate the analysis, we make the following assumptions.

Assumption 1 (Lipschitz-Continuous Gradient): The objective function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth, i.e. for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, it is satisfied

$$f(\mathbf{w}) \leq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle + \frac{L}{2} \|\mathbf{w} - \mathbf{v}\|^2 \quad (12)$$

where L is a positive constant.

Assumption 2 (SGD Sampling Noise): For every client n , the stochastic gradient $\nabla \tilde{f}_n(\mathbf{w}; \mathcal{B}_n) = \frac{1}{B} \sum_{i \in \mathcal{B}_n} \nabla \ell(\mathbf{w}; i)$, calculated based on an independent mini-batch \mathcal{B}_n containing B data samples, is an unbiased estimation of $\nabla f_n(\mathbf{w}; \mathcal{B}_n)$ with bounded variance, i.e.,

$$\mathbb{E} \left[\nabla \tilde{f}_n(\mathbf{w}; \mathcal{B}_n) \right] = \nabla f_n(\mathbf{w}; D_n), \quad (13)$$

$$\mathbb{E} \left[\left\| \nabla \tilde{f}_n(\mathbf{w}; \mathcal{B}_n) - \nabla f_n(\mathbf{w}; D_n) \right\|^2 \right] \leq \frac{\sigma_s^2}{B}. \quad (14)$$

Assumption 3 (Gradient Bound): The gradient of functions $f_n(\mathbf{w})$ is bounded, i.e., for $\forall n \in [N]$, there exists a positive constant G that

$$\|\nabla f_n(\mathbf{w})\| \leq G. \quad (15)$$

Notably, Assumptions 1 and 2 are commonly adopted in previous works [5], [9], [15], while Assumption 3 generally holds in the setting of over-the-air computations due to the stipulation of maximum transmit power.¹ It is important to mention that we only assume the smoothness of the loss function, making it applicable to even settings involving (deep) neural networks with non-convex objective functions.

We start by investigating the effect of channel fading on the aggregated global gradient. By the averaging operation to the globally aggregated gradient, we establish the following concentration inequality.

Lemma 1: Let $\nu > 0$ be a constant, and denote by $\beta_\nu = \mathbb{P}(|c_{n,t} - \mu_c| > \nu)$. At any communication round t , for the i -th entry of the aggregated gradient, $\forall i \in [d]$, the following holds for all $\varepsilon \geq 0$

$$\begin{aligned} & \mathbb{P} \left(\frac{1}{N} \sum_{n=1}^N c_{n,t} \nabla f_{n,(i)}(\mathbf{w}_t) - \frac{1}{N} \sum_{n=1}^N \mu_c \nabla f_{n,(i)}(\mathbf{w}_t) \geq \varepsilon + 2\nu \beta_\nu^N G \right) \\ & \leq \beta_\nu^N + \exp \left(-N \frac{\varepsilon^2}{2\nu^2 G^2} \right). \end{aligned} \quad (16)$$

Proof: Please see Appendix B. \square

Remark 3: When $N \gg 1$, the right-hand-side of (16) approaches zero (by controlling ν we can ensure $\beta_\nu \ll 1$). This resembles the channel hardening effect [20], whereby the impairment of random channel fading is averaged out in the presence of massive clients (which is the adequate regime for employing over-the-air model aggregation).

Following Lemma 1, we can tightly approximate the global gradient in (6) as follows:

$$\mathbf{g}_t \approx \frac{1}{N} \sum_{n=1}^N \mu_c \nabla \tilde{f}_n(\mathbf{w}_t) + \frac{1}{N} \boldsymbol{\xi}_t. \quad (17)$$

We are now equipped to present the convergence rate. The result under a fixed learning rate is given as follows.

¹Due to the constraint of power budget at the radio front end, if some entries of the upload parameter are excessively large, they need to be trimmed before being modulated to the radio signal and sent out.

Theorem 2: Under the employed edge learning system, if the global and local learning rates are set as $\eta_t = \frac{1}{\mu_c} \eta_l = \frac{1}{L\mu_c}$, then the model training converges as

$$R_N^{(T)} \leq \frac{2L(f(\mathbf{w}_0) - f(\mathbf{w}^*))}{TE} + \frac{d\sigma_z^2}{\mu_c^2 N^2 E} + \frac{\sigma_s^2}{NBE} + G^2 \frac{(E-1)(2E+5)}{6}. \quad (18)$$

Proof: Please see Appendix C. \square

Our analysis emphasizes the impact of the number of clients N on the convergence rate, particularly its effect on local SGD estimation noise and channel noise. Correspondingly, we can derive the convergence rate under a decaying learning rate.

Corollary 1: Under the employed edge learning system, if the global and local learning rates are set as $\eta_t = \frac{1}{\mu_c} \eta_l = \frac{\eta_0}{1+t}$, where $\eta_0 < \frac{1}{L\mu_c E}$, then the model training converges as

$$\begin{aligned} & \min_{t=0,1,\dots,T-1} \mathbb{E} \left[\left\| \nabla f(\mathbf{w}_t) \right\|^2 \right] \leq \frac{2(f(\mathbf{w}_0) - f(\mathbf{w}^*))}{\mu_c \eta_0 E \log T} + \frac{L\mu_c \eta_0 \pi^2}{3 \log T} \\ & \times \left(\frac{d\sigma_z^2}{\mu_c^2 N^2 E} + \frac{\sigma_s^2}{NB} + \frac{G^2 L^2 \eta_0^2 \mu_c^2 \pi^2 E(E-1)(2E-1)}{90} \right). \end{aligned} \quad (19)$$

Proof: The proof is similar to that for Theorem 1 and thus omitted. \square

Remark 4: Under a fixed learning rate, the global model ultimately lands within a noisy ball around optimality at a rate of $\mathcal{O}(\frac{1}{T})$. In contrast, by adopting a decaying learning rate, the model can arrive at the minima, with a convergence rate at the order of $\mathcal{O}(\frac{1}{\log(T)})$.

Remark 5: If we denote by local epoch E as a portion of the mini-batch counts, i.e., $E = \tau \frac{M}{B}$, we can rewrite the error bound as $R_N^{(T)} \leq \frac{2L(f(\mathbf{w}_0) - f(\mathbf{w}^*))}{\tau MT} B + \frac{d\sigma_z^2}{\tau M \mu_c^2 N^2} B + \frac{\sigma_s^2}{\tau MN} + G^2 \frac{(\frac{\tau M}{B} - 1)(2\frac{\tau M}{B} + 5)}{6}$. This implies that properly tuning the batch size helps balance computational efficiency and model accuracy, reducing residual errors during training.

C. Discussions

Using the above analysis, we demonstrate three notable benefits of system scaling up.

1) *Enhancing Privacy Protection:* When $N \rightarrow \infty$, Theorem 1 yields

$$I_N^{(t)} \leq \frac{C_g d^*}{N-1} + \frac{1}{2} \sum_{i=1}^{d^*} \frac{\lambda_{t,i}}{(N-1)\lambda_{t,i} + \sigma_z^2} \sim \mathcal{O}\left(\frac{1}{N}\right). \quad (20)$$

This indicates that the privacy leakage of the noisy aggregated gradient decays at the rate of $\mathcal{O}(1/N)$. In essence, a large number of participants facilitates each individual to hide its information in the crowd, thus improving privacy protection.

2) *Mitigating Channel Impairments:* From Lemma 1, we observe that the difference between the desired global gradient $\frac{1}{N} \sum_{n=1}^N \nabla f_n(\mathbf{w}_t)$ and its perturbed version $\frac{1}{N} \sum_{n=1}^N \frac{c_{n,t}}{\mu_c} \nabla f_n(\mathbf{w}_t)$ reduces at the rate of $\mathcal{O}(\exp(-N))$, namely, the channel hardening effect (quickly) becomes evident when the number of clients increases. As a result, the

need for precise channel estimation and power control to counteract channel fading is greatly alleviated. Moreover, while [21] demonstrates that the fading distortions in over-the-air federated learning diminish as the number of antennas grows, our analysis reveals that with a large number of clients present, the averaging automatically ignites channel hardening, which alleviates channel impairments in over-the-air computations.

3) *Improving Training Efficiency*: In addition to the channel hardening effect, a large number of participating clients is also instrumental in reducing the estimation and communication noise, which, in turn, improves training efficiency. To be more concrete, let us take $E = 1$ as an example (also known as the FedSGD). In this case, the estimation error at global iteration T can be bounded by the following

$$R_N^{(T)} \leq \frac{2L(f(\mathbf{w}_0) - f(\mathbf{w}^*))}{T} + \frac{d\sigma_z^2}{\mu_c^2 N^2} + \frac{\sigma_s^2}{NB}. \quad (21)$$

This result indicates that increasing N not only reduces the impact of thermal noise but, more importantly, also decreases the estimation noise. Indeed, with $N \rightarrow \infty$, the residual error (i.e. the second and third terms on the right-hand-side of (21)) vanishes, implying that the stochastic gradient obtained from an analog, noisy transmission would be equivalent to a global gradient achieved from a noiseless transmission.

Our analytical derivations demonstrate that increasing the number of clients in the system enhances both privacy protection and training efficiency. As illustrated in the next section, these findings are further validated through simulations.

IV. EXPERIMENTAL RESULTS

A. Setup

We evaluate the performance of the edge learning system by experiments on the CIFAR-10 [22] and EMNIST [23] dataset using neural network model architectures including the ResNet-18 [24] and a LeNet-5 like Convolutional Neural Network (CNN) model [25], respectively. We vary the number of clients while keeping the dataset size per client fixed.

Unless otherwise stated, we use the Rayleigh fading to model the channel gain with an average of $\mu_c = 1$. We further consider the same learning rate $\eta = 0.03$ for both local and global training, i.i.d. data set, local epoch $E = \frac{M}{B}$, and local batch size $B = 50$. All experiments are implemented with Pytorch on NVIDIA RTX 3090 GPU.

We use the Mutual Information Neural Estimator (MINE) [26] to estimate the MI between each client's local gradient $\nabla f_n(\mathbf{w}_t)$ and the aggregated global gradient \mathbf{g}_t . We follow similar procedures in [19] to obtain samples in the context of over-the-air computations. Specifically, we employ a fully connected neural network with two hidden layers, each containing 256 neurons, and using a learning rate of $\eta = 10^{-4}$ over 1000 training iterations. For the noisy setting, we sample M sets of noise parameters $\{(\{c_{n,t}^{(m)}\}_{n=1}^N; \xi_t^{(m)})\}_{m=1}^M$, and then loop these samples to calculate the aggregated model update for each gradient sample. This process yields M MI estimates for a single communication round, with the average serving as the

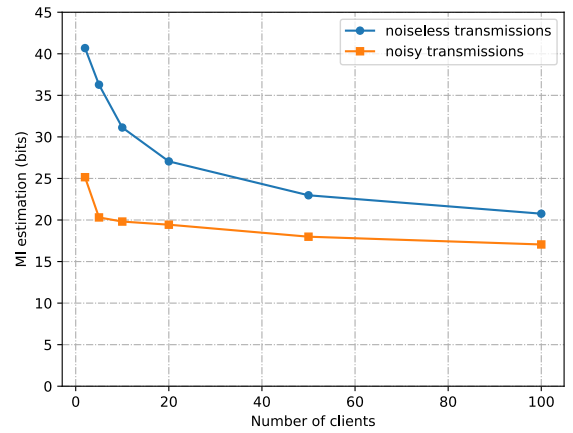


Fig. 2: Impact of client number N on privacy leakage, evaluated from training a CNN on the EMNIST dataset.

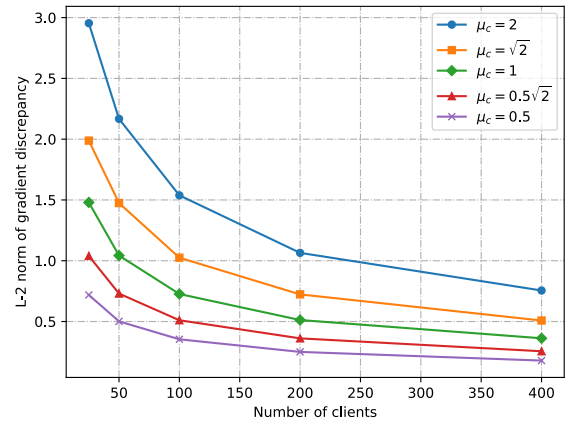
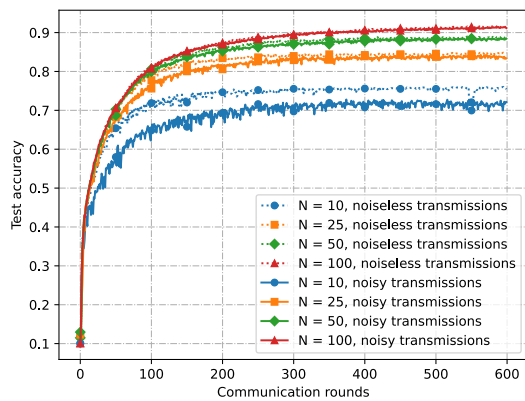


Fig. 3: Visualizing the channel hardening effect, exemplified by training a CNN on the EMNIST dataset.

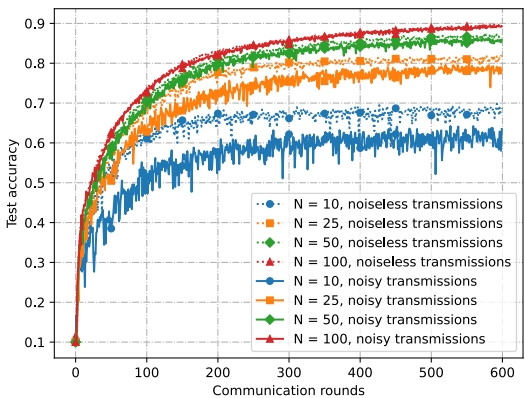
final result. We evaluate the MI for each network layer and report the total MI across all layers as the final result.

B. Performance Evaluation

Fig. 2 draws the MI estimation as a function of the client number under noisy and noiseless transmission scenarios. From this figure, we can see that as the number of clients increases, the privacy leakage of each participant decreases, which is consistent with our analysis in Theorem 1. As such, it validates the benefits of system scaling up in enhancing privacy protection. Notably, given the same client number, the noisy setting results in a smaller MI compared to the noiseless transmission scenario. This observation confirms the conclusion in Remark 2 that channel distortion offers inherent privacy protection for over-the-air computing. The main reason can be attributed to the fact that the superposition property of multiple access channels effectively obscures individual client information within the globally aggregated gradient, whereas channel noise further perturbs the result. Moreover, as the number of clients increases, the challenge of distinguishing an individual local gradient from the aggregated gradient grows significantly, rendering it akin to finding a needle in a haystack.



(a) i.i.d. data set



(b) non-i.i.d. data set

Fig. 4: Impact of the number of clients N on test accuracy, exemplified by training ResNet-18 on the CIFAR-10 dataset.

In Fig. 3, we verify the channel hardening effect observed in Lemma 1. We use the actual channel coefficients $\{c_{n,t}\}_{n=1}^N$ and their mean μ_c to calculate a pair of aggregated gradients $\{\frac{1}{N} \sum_{n=1}^N c_{n,t} \nabla f_n(\mathbf{w}_t); \frac{1}{N} \sum_{n=1}^N \mu_c \nabla f_n(\mathbf{w}_t)\}$ at each communication round t . We then compute the L-2 norm of the discrepancy of the aggregated gradients pair and report the average across 200 global communication rounds. The figure reveals that an increase in the number of clients leads to a decrease in the gradient discrepancy. This confirms that scaling up the system mitigates the fluctuations in small-scale fading, driving the noisy global gradient close to its unperturbed version, which effectively enhances the system’s robustness.

We investigate the effect of system scaling up on the test accuracy under both i.i.d. and non-i.i.d. data sets in Fig. 4a and Fig. 4b, respectively. The non-i.i.d. data partitions are implemented via the widely adopted symmetric Dirichlet distribution [27]. The results show that test accuracy consistently improves with an increasing number of clients, demonstrating a positive influence from the enlarged scale of the system. Another noteworthy observation is that the convergence curve becomes smoother and approaches the behavior observed under noiseless transmission as the number of clients increases, whereas it exhibits more fluctuations with fewer clients. These findings highlight the dual benefits of increasing the number of

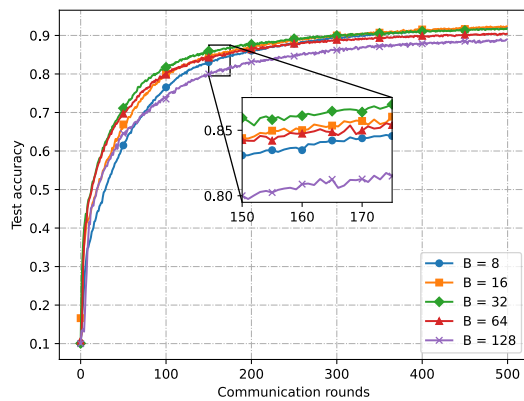


Fig. 5: Impact of batch size B on test accuracy, exemplified by training ResNet-18 on the CIFAR-10 dataset.

participating clients: It enhances the training performance by providing richer training datasets and reducing the gradient estimation and communication noise in model updates, thus leading to a more stable and swift convergence process.

In Fig. 5, we examine the effect of local batch size with $N = 100$ clients. We observe that as the batch size increases, the convergence rate initially improves but eventually slows down. The reason for this is twofold: (a) a larger batch size reduces the SGD sampling noise, resulting in more stable updates and (b) a smaller batch size facilitates more frequent updates to traverse the local dataset. This suggests the existence of an optimal batch size for a fixed learning rate and local dataset size that accelerates convergence.

V. CONCLUSION

We have conducted a theoretical study of analog over-the-air model training. We focus mainly on the system scale-up and evaluate its impact on privacy protection and training efficiency. Specifically, we have established an analytical expression for assessing privacy leakage regarding mutual information during each round of global communication. We have also developed a Bernstein-like inequality that unveils a channel hardening effect in analog over-the-air model aggregation. Subsequently, we have derived the convergence rate for non-convex loss functions in federated learning over the air. Our analysis reveals three benefits conferred by scaling up the system, namely, with an increasing number of clients: (i) the privacy leakage is substantially reduced since every client’s gradient is obscured in the globally aggregated model, (ii) the channel hardening effect becomes more evident, eliminating the impairments of small-scale fading, and (iii) the convergence rate can be accelerated as the thermal noise and gradient estimation noise can be further decreased. We have validated these theoretical findings through various experiments.

VI. APPENDIX

A. Proof of Theorem 1

Without loss of generality, we derive the upper bound for the following term (as change to the others only requires

undergoing a permutation of client indices)

$$I(\nabla \tilde{f}_N(\mathbf{w}_t); \mathbf{g}_t | \{\mathbf{g}_p\}_{p \in [t-1]}). \quad (22)$$

Similar to [19], we define $\nabla \tilde{f}_n(\mathbf{w}_t) \in \mathbb{R}^{d^*}$, where $d^* \leq d$, which is a sub-vector with a rank- d^* covariance matrix, induced from $\nabla f_n(\mathbf{w}_t)$. For all $n \in [N]$, $\nabla \tilde{f}_n(\mathbf{w}_t)$ has a non-singular covariance matrix \mathbf{K}_t with mean $\boldsymbol{\mu}_t$. We also define $\mathbf{F}_N^{(t)} = \frac{1}{\sqrt{N}} \left(\sum_{n=1}^N c_{n,t} \nabla \tilde{f}_n(\mathbf{w}_t) + \tilde{\boldsymbol{\xi}}_t \right)$. Then, we have

$$\begin{aligned} & I(\nabla \tilde{f}_N(\mathbf{w}_t); \mathbf{g}_t | \{\mathbf{g}_p\}_{p \in [t-1]}) \stackrel{(a)}{=} I(\nabla \tilde{f}_N(\mathbf{w}_t); \mathbf{F}_N^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) \\ & = h(\nabla \tilde{f}_N(\mathbf{w}_t) | \{\mathbf{g}_p\}_{p \in [t-1]}) + h(\mathbf{F}_N^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) \\ & \quad - h \left(\begin{bmatrix} \mathbf{I}_{d^*} & \mathbf{0}_{d^*} \\ \frac{c_{n,t}}{\sqrt{N}} \mathbf{I}_{d^*} & \frac{\sqrt{N-1}}{\sqrt{N}} \mathbf{I}_{d^*} \end{bmatrix} \left[\begin{array}{c} \nabla \tilde{f}_N(\mathbf{w}_t) \\ \mathbf{F}_{N-1}^{(t)} \end{array} \right] \middle| \{\mathbf{g}_p\}_{p \in [t-1]} \right) \\ & \stackrel{(b)}{=} h(\nabla \tilde{f}_N(\mathbf{w}_t) | \{\mathbf{g}_p\}_{p \in [t-1]}) + h(\mathbf{F}_N^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) \\ & \quad - h(\nabla \tilde{f}_N(\mathbf{w}_t) | \{\mathbf{g}_p\}_{p \in [t-1]}) - h(\mathbf{F}_{N-1}^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) \\ & \quad - \log \left| \det \begin{bmatrix} \mathbf{I}_{d^*} & \mathbf{0}_{d^*} \\ \frac{c_{n,t}}{\sqrt{N}} \mathbf{I}_{d^*} & \frac{\sqrt{N-1}}{\sqrt{N}} \mathbf{I}_{d^*} \end{bmatrix} \right| \\ & = h(\mathbf{F}_N^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) - h(\mathbf{F}_{N-1}^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) + \frac{d^*}{2} \log \left(\frac{N}{N-1} \right) \end{aligned} \quad (23)$$

where (a) holds because MI is invariant under multiplying with a constant, and (b) follows from the property of the entropy of linear transformation of random vectors [28].

To further bound the first two terms in the last equality of (23), we denote by $\mathbf{s}_{n,t} = c_{n,t} \nabla \tilde{f}_n(\mathbf{w}_t)$, which has the covariance matrix $\boldsymbol{\Sigma}_t = (\mu_c^2 + \sigma_c^2) \mathbf{K}_t + \sigma_c^2 \boldsymbol{\mu}_t \boldsymbol{\mu}_t^\top$. Leveraging Definition 1, we have $\hat{\mathbf{s}}_{n,t} = \boldsymbol{\Sigma}_t^{-1/2} (\mathbf{s}_{n,t} - \mu_c \boldsymbol{\mu}_t)$ with zero mean and identity covariance. Then, we characterize the following entropy term for any $M \in \mathbb{N}$

$$\begin{aligned} & h(\mathbf{F}_M^{(t)} | \{\mathbf{g}_p\}_{p \in [t-1]}) = h \left(\frac{1}{\sqrt{M}} \sum_{n=1}^M \mathbf{s}_{n,t} \middle| \{\mathbf{g}_p\}_{p \in [t-1]} \right) \\ & = h \left(\frac{\boldsymbol{\Sigma}_t^{1/2}}{\sqrt{M}} \left(\sum_{n=1}^M \hat{\mathbf{s}}_{n,t} + \boldsymbol{\Sigma}_t^{-1/2} \tilde{\boldsymbol{\xi}}_t \right) \middle| \{\mathbf{g}_p\}_{p \in [t-1]} \right) \\ & = \log \left| \det \boldsymbol{\Sigma}_t^{1/2} \right| + \underbrace{h \left(\frac{1}{\sqrt{M}} \left(\sum_{n=1}^M \hat{\mathbf{s}}_{n,t} + \boldsymbol{\Sigma}_t^{-1/2} \tilde{\boldsymbol{\xi}}_t \right) \middle| \{\mathbf{g}_p\}_{p \in [t-1]} \right)}_{H_M}. \end{aligned} \quad (24)$$

The upper bound of H_M can be obtained by matching the first and second moment to a Gaussian-distributed vector, as follows:

$$\begin{aligned} H_M & \leq \frac{1}{2} \log \left[(2\pi e)^{d^*} \det \left(\mathbf{I}_{d^*} + \frac{1}{M} \sigma_z^2 \boldsymbol{\Sigma}_t^{-1} \right) \right] \\ & = \frac{1}{2} \sum_{i=1}^{d^*} \log \left[2\pi e \left(1 + \frac{\sigma_z^2}{M \lambda_{t,i}} \right) \right] \end{aligned} \quad (25)$$

where $\{\lambda_{t,i}\}_{i=1}^{d^*}$ represent the eigenvalues of $\boldsymbol{\Sigma}_t$. And a lower bound of H_M follows from leveraging the Berry-Esseen style bounds for the entropic central limit theorem [29]:

$$\begin{aligned} H_M & = \sum_{i=1}^{d^*} h \left(\frac{1}{\sqrt{M}} \left(\sum_{n=1}^M \hat{\mathbf{s}}_{n,t,(i)} + (\boldsymbol{\Sigma}_t^{-1/2} \tilde{\boldsymbol{\xi}}_t)_{(i)} \right) \middle| \{\mathbf{g}_p\}_{p \in [t-1]} \right) \\ & \geq \frac{1}{2} \sum_{i=1}^{d^*} \log \left[2\pi e \left(1 + \frac{\sigma_z^2}{M \lambda_{t,i}} \right) \right] - \frac{d^* C_{\bar{g}}}{M} \end{aligned} \quad (26)$$

where $C_{\bar{g}}$ denotes a specific constant associated with the finite fourth moment of $\hat{\mathbf{s}}_{n,t,(i)}$.

Finally, by substituting (24) into (23), we have

$$I(\nabla \tilde{f}_N(\mathbf{w}_t); \mathbf{g}_t | \{\mathbf{g}_p\}_{p \in [t-1]}) = H_N - H_{N-1} + \frac{d^*}{2} \log \left(\frac{N}{N-1} \right). \quad (27)$$

The proof is completed by substituting (25) (with $M = N$) and (26) (with $M = N - 1$) into (27).

B. Proof of Lemma 1

We define $\mathbf{z}_{n,t} = (c_{n,t} - \mu_c) \nabla f_n(\mathbf{w}_t)$ for each $n \in [N]$. Since $\{c_{n,t}\}_{n=1}^N$ are i.i.d., it follows that $\mathbb{E}[\mathbf{z}_{n,t}] = \mathbf{0}$ and $\mathbb{E}[\|\mathbf{z}_{n,t,(i)}\|^2] \leq \mathbb{E}[\|\mathbf{z}_{n,t}\|^2] \leq \sigma_c^2 G^2$. Within this regime, we apply the extension of the vector Bernstein inequality [30] and obtain the following for $0 < \varepsilon < \sigma_c G$:

$$\mathbb{P} \left(\left\| \frac{1}{N} \sum_{n=1}^N \mathbf{z}_{n,t,(i)} \right\| \geq \varepsilon \right) \leq \exp \left(-N \frac{\varepsilon^2}{8\sigma_c^2 G^2} + \frac{1}{4} \right). \quad (28)$$

To handle the remaining regime, we invoke Mcdiarmid's inequality [31]. Specifically, we define a function $F_{(i)}(\mathbf{c}_t) = \frac{1}{N} \sum_{n=1}^N c_{n,t} \nabla f_{n,(i)}(\mathbf{w}_t)$, $\forall i \in [d]$, with $\mathbf{c}_t = (c_{1,t}, \dots, c_{N,t})$. Given $\beta_\nu = \mathbb{P}(|c_{n,t} - \mu_c| > \nu)$, let $\mathcal{Y} \subset (\mathbb{R}^+)^N$ be the subset where each entry of $\mathbf{c}_t \in \mathcal{Y}$ satisfies $|c_{n,t} - \mu_c| \leq \nu$. Thus $F_{(i)}(\mathbf{c}_t)$ has a $\frac{2\nu G}{N}$ -bounded difference [31] over \mathcal{Y} , and $\mathbb{P}(\mathbf{c}_t \notin \mathcal{Y}) = \beta_\nu^N$. Therefore, for all $\varepsilon \geq 0$, we have

$$\mathbb{P}(F_{(i)}(\mathbf{c}_t) - \mathbb{E}[F_{(i)}(\mathbf{c}_t) | \mathbf{c}_t \in \mathcal{Y}] \geq \varepsilon + 2\nu \beta_\nu^N G) \leq \beta_\nu^N + \exp \left(-\frac{N \varepsilon^2}{2\nu^2 G^2} \right). \quad (29)$$

C. Proof of Theorem 2

For ease of exposition, let us define $\tilde{\mathbf{g}}_t = \frac{1}{\mu_c E} \mathbf{g}_t$. And we denote by $\nabla \tilde{f}_{n,t}^{(k)} = \frac{1}{B} \sum_{j=1}^B \nabla \ell(\mathbf{w}_{n,t}^{(k)}; \theta_{n,t}^{(k)}(j))$ and $\nabla f_{n,t}^{(k)} = \frac{1}{M} \sum_{j=1}^M \nabla \ell(\mathbf{w}_{n,t}^{(k)}; \theta_{n,t}^{(k)}(j))$. Then, using the smoothness of $f(\mathbf{w})$, we have

$$\begin{aligned} \mathbb{E}[f(\mathbf{w}_{t+1})] - \mathbb{E}[f(\mathbf{w}_t)] & \leq \mathbb{E}[\langle \nabla f(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle] + \frac{L}{2} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\ & = -\frac{\eta_t E \mu_c}{2} \mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2] - \frac{\eta_t E \mu_c (1 - \eta_t E \mu_c L)}{2} \mathbb{E}[\|\tilde{\mathbf{g}}_t\|^2] \\ & \quad + \frac{\eta_t E \mu_c}{2} \mathbb{E}[\|\nabla f(\mathbf{w}_t) - \tilde{\mathbf{g}}_t\|^2]. \end{aligned} \quad (30)$$

By noticing that $\mathbb{E}[\tilde{\boldsymbol{\xi}}_t] = \mathbf{0}$ and applying Assumption 3, we can bound $\mathbb{E}[\|\tilde{\mathbf{g}}_t\|^2]$ as follows:

$$\begin{aligned} \mathbb{E}[\|\tilde{\mathbf{g}}_t\|^2] & = \mathbb{E} \left[\left\| \frac{1}{E} \left(\frac{1}{N} \sum_{n=1}^N \sum_{k=0}^{E-1} \nabla \tilde{f}_{n,t}^{(k)} + \frac{\boldsymbol{\xi}_t}{\mu_c N} \right) \right\|^2 \right] \\ & = \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \frac{1}{E} \sum_{k=0}^{E-1} \nabla \tilde{f}_{n,t}^{(k)} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{\boldsymbol{\xi}_t}{\mu_c E N} \right\|^2 \right] \leq G^2 + \frac{d \sigma_z^2}{\mu_c^2 E^2 N^2}. \end{aligned} \quad (31)$$

Next, we expand $\mathbb{E}[\|\nabla f(\mathbf{w}_t) - \tilde{\mathbf{g}}_t\|^2]$ as follows:

$$\begin{aligned} & \mathbb{E}[\|\nabla f(\mathbf{w}_t) - \tilde{\mathbf{g}}_t\|^2] \\ & = \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \nabla f_{n,t}^{(0)} - \frac{1}{E} \left(\frac{1}{N} \sum_{n=1}^N \sum_{k=0}^{E-1} \nabla \tilde{f}_{n,t}^{(k)} + \frac{\boldsymbol{\xi}_t}{\mu_c N} \right) \right\|^2 \right] \\ & = \mathbb{E} \left[\underbrace{\left\| \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{E} \sum_{k=0}^{E-1} \nabla \tilde{f}_{n,t}^{(k)} - \nabla f_{n,t}^{(0)} \right) \right\|^2}_{e_{n,t}} \right] + \mathbb{E} \left[\left\| \frac{\boldsymbol{\xi}_t}{\mu_c E N} \right\|^2 \right] \end{aligned} \quad (32)$$

where $e_{n,t}$ can be further decomposed into the following:

$$e_{n,t} = \underbrace{\frac{1}{E} \sum_{k=0}^{E-1} (\nabla \tilde{f}_{n,t}^{(k)} - \nabla f_{n,t}^{(k)})}_{\hat{e}_{n,t}} + \underbrace{\frac{1}{E} \sum_{k=0}^{E-1} (\nabla f_{n,t}^{(k)} - \nabla f_{n,t}^{(0)})}_{\bar{e}_{n,t}}. \quad (33)$$

Subsequently, we bound $\hat{e}_{n,t}$ and $\bar{e}_{n,t}$, respectively. According to Assumption 2, we have $\mathbb{E}[\hat{e}_{n,t}] = 0$, hence

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \hat{e}_{n,t} \right\|^2 \right] = \frac{1}{N^2} \sum_{n=1}^N \frac{1}{E^2} \sum_{k=0}^{E-1} \mathbb{E} \left[\left\| \nabla \tilde{f}_{n,t}^{(k)} - \nabla f_{n,t}^{(k)} \right\|^2 \right] \leq \frac{\sigma_s^2}{NBE}. \quad (34)$$

On the other hand, using Assumption 3 and smoothness of the loss function, we can bound $\bar{e}_{n,t}$ via the following:

$$\begin{aligned} \mathbb{E} \left[\left\| \bar{e}_{n,t} \right\|^2 \right] &\leq \frac{1}{E} \sum_{k=0}^{E-1} \mathbb{E} \left[\left\| \nabla f_{n,t}^{(k)} - \nabla f_{n,t}^{(0)} \right\|^2 \right] \leq \frac{L^2}{E} \sum_{k=0}^{E-1} \mathbb{E} \left[\left\| \mathbf{w}_{n,t}^{(k)} - \mathbf{w}_t \right\|^2 \right] \\ &= \frac{L^2}{E} \sum_{k=0}^{E-1} \mathbb{E} \left[\left\| \eta_l \sum_{p=0}^{k-1} \nabla \tilde{f}_{n,t}^{(p)} \right\|^2 \right] \leq \frac{L^2 \eta_l^2 G^2 (E-1)(2E-1)}{6}. \end{aligned} \quad (35)$$

To this end, we obtain

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla f(\mathbf{w}_t) - \tilde{\mathbf{g}}_t \right\|^2 \right] &= \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N e_{n,t} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{\boldsymbol{\xi}_t}{\mu_c EN} \right\|^2 \right] \\ &= \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \hat{e}_{n,t} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \bar{e}_{n,t} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{\boldsymbol{\xi}_t}{\mu_c EN} \right\|^2 \right] \\ &\leq \frac{\sigma_s^2}{NBE} + \frac{L^2 \eta_l^2 G^2 (E-1)(2E-1)}{6} + \frac{d\sigma_z^2}{\mu_c^2 E^2 N^2}. \end{aligned} \quad (36)$$

Since $\eta_t = \frac{1}{\mu_c} \eta_l = \frac{1}{\mu_c L}$, by simple algebra, we have

$$\begin{aligned} \mathbb{E}[f(\mathbf{w}_{t+1})] - f(\mathbf{w}^*) &\leq \mathbb{E}[f(\mathbf{w}_t)] - f(\mathbf{w}^*) - \frac{E}{2L} \mathbb{E} \left[\left\| \nabla f(\mathbf{w}_t) \right\|^2 \right] \\ &\quad + \frac{d\sigma_z^2}{2L\mu_c^2 N^2} + \frac{\sigma_s^2}{2NBL} + \frac{E(E-1)(2E+5)G^2}{12L}. \end{aligned} \quad (37)$$

By summing up (37) for $t = 0, 1, \dots, T-1$ and rearranging the above formula, we complete the proof.

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 *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, Apr. 2017, pp. 1273–1282.
- [2] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, Jun. 2020.
- [3] 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, Jan. 2020.
- [4] 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, Mar. 2020.
- [5] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.
- [6] 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. Topics Signal Process.*, vol. 16, no. 3, pp. 406–419, Apr. 2022.
- [7] H. H. Yang, Z. Chen, and T. Q. Quek, "Unleashing edgeless federated learning with analog transmissions," *IEEE Trans. Signal Process.*, vol. 72, pp. 774–791, Jan. 2024.
- [8] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, Aug. 2020.
- [9] 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, Jan. 2022.
- [10] W. Guo, R. Li, C. Huang, X. Qin, K. Shen, and W. Zhang, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, Aug. 2022.
- [11] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019.
- [12] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients—How easy is it to break privacy in federated learning?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020, pp. 16937–16947.
- [13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security and Priv. (SP)*, San Jose, CA, May 2017, pp. 3–18.
- [14] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–6.
- [15] 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. 2021.
- [16] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5194–5208, Aug. 2021.
- [17] X. Li, "RSS-based location estimation with unknown pathloss model," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3626–3633, Dec. 2006.
- [18] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, Dallas, TX, Oct. 2017, pp. 1175–1191.
- [19] A. R. Elkordy, J. Zhang, Y. H. Ezzeldin, K. Psounis, and S. Avestimehr, "How much privacy does federated learning with secure aggregation guarantee?" in *Proc. Priv. Enhanc. Technol. (PoPETs)*, Lausanne, Switzerland, Jul. 2023.
- [20] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sept. 2004.
- [21] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, Aug. 2021.
- [22] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 4*, 2009.
- [23] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, Anchorage, AK, May 2017, pp. 2921–2926.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 770–778.
- [25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [26] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 531–540.
- [27] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *Available as ArXiv:1909.06335*, 2019.
- [28] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, USA, 2006.
- [29] S. G. Bobkov, G. P. Chistyakov, and F. Götze, "Berry-eseen bounds in the entropic central limit theorem," *Probab. Theory Relat. Fields*, vol. 159, no. 3, pp. 435–478, Aug. 2014.
- [30] J. M. Kohler and A. Lucchi, "Sub-sampled cubic regularization for non-convex optimization," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 1895–1904.
- [31] R. Combes, "An extension of McDiarmid's inequality," *Available as ArXiv:1511.05240*, 2015.