Privacy For Free: Wireless Federated Learning Via Uncoded Transmission With Adaptive Power Control

Dongzhu Liu and Osvaldo Simeone

***Abstract*—Federated Learning (FL) refers to distributed protocols that avoid direct raw data exchange among the participating devices while training for a common learning task. This way, FL can potentially reduce the information on the local data sets that is leaked via communications. In order to provide formal privacy guarantees, however, it is generally necessary to put in place additional masking mechanisms. When FL is implemented in wireless systems via uncoded transmission, the channel noise can directly act as a privacy-inducing mechanism. This paper demonstrates that, as long as the privacy constraint level, measured via differential privacy (DP), is below a threshold that decreases with the signal-to-noise ratio (SNR), uncoded transmission achieves privacy "for free", i.e., without affecting the learning performance. More generally, this work studies adaptive power allocation (PA) for decentralized gradient descent in wireless FL with the aim of minimizing the learning optimality gap under privacy and power constraints. Both orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) transmission with "over-the-air-computing" are studied, and solutions are obtained in closed form for an offline optimization setting. Furthermore, heuristic online methods are proposed that leverage iterative one-step-ahead optimization. The importance of dynamic PA and the potential benefits of NOMA versus OMA are demonstrated through extensive simulations.**

***Index Terms*—Federated learning, differential privacy, adaptive power control, uncoded transmission.**

I. Introduction

In modern wireless systems, mobile devices generate and store data that can be utilized to train machine learning models [1]–[3]. While data at one device may be insufficient to obtain effective trained solutions, networked devices can benefit from data stored at other devices via communications. Federated learning (FL) refers to decentralized training protocols that avoid direct data sharing among devices, while exchanging information about the local models [4], [5]. This has the potential benefits of reducing the communication load and of leaking less information about the local data sets at the devices [6]–[8]. A well-established measure the privacy of local data sets with respect to disclosed aggregate statistics is differential privacy (DP) [9]. Typical DP mechanisms randomize the disclosed statistics by adding random noise [9]. This creates a trade-off between accuracy and privacy, as determined by the amount of added noise.

This paper investigates the idea of letting the channel noise serve as privacy mechanism. To this end, we focus on uncoded transmission of the gradients using either orthogonal or non-orthogonal protocols, and we analytically demonstrate that for these transmission schemes, privacy may be obtained "for free". This is in the sense that enforcing a DP constraint causes no performance loss with respect to a non-private design as long as the signal-to-noise ratio (SNR) is sufficiently low. More generally, we introduce a novel optimal closed-form adaptive transmit power control strategy that optimizes the learning performance while ensuring DP requirements.

The authors are with King's Communications, Learning, and Information Processing (KCLIP) lab at the Department of Engineering of Kings College London, UK (emails: dongzhu.liu@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). The authors have received funding from the European Research Council (ERC) under the European Unions Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

A. Wireless Federated Learning

A illustrated in Fig. 1, typical wireless FL protocols iterate between local adaptation and centralized combining. Adaptation involves local optimization steps based on a device's data, while central aggregation amounts to averaging operations on the devices' updates. To reduce the time to convergence, recent work has proposed to leverage computations over multiple access channels [10] as a primitive for the global combining step [11]–[15]. Accordingly, all devices simultaneously transmit their updates to edge server using uncoded transmission, which are aggregated "over-the-air" by exploiting the waveform-superposition property of a multi-access channel.

To further improve the bandwidth efficiency of this non-orthogonal multiple access (NOMA) scheme, devices can pre-process the analog updates via sparsification based dimensionality reduction [13]. And the learning performance of this approach can be further enhanced by gradient aware power control [14] and joint device selection and beamforming design [15].

As a more conventional solution, FL can be implemented using digital coded transmission. Under digital coded transmission and orthogonal multiple access (OMA), quantization of the local gradients has been proposed to trade off communication bandwidth and convergence rate [16], with each component of the gradient being represented even by a single bit [17]. More complex quantization schemes include hierarchical quantization via a low-dimensional codebook on a Grassmann manifold [18]. With NOMA, reference [19] proposes a strategy whereby each device quantizes the gradient based on its informativeness and on the channel condition. An alternative is to use one-bit quantization followed by BPSK/QPSK modulation at the devices with NOMA, and to estimate the aggregated gradient at the edger sever using majority voting [20].

B. Differential Privacy for Federated Learning

According to its original motivation, FL may have desirable privacy properties since training is conducted in a distributed

manner without sharing the raw data. Nevertheless, the model updates shared by the devices may reveal information about local data. For example, a malicious server could potentially infer the presence of an individual data sample from a learnt model by membership inference attack [21] or model inversion attack [7]. DP quantifies information leaked about individual data points by measuring the sensitivity of the disclosed statistics to changes in the input data set at a single data point. DP can be guaranteed by introducing a level of uncertainty into the released model that is sufficient to mask the contribution of any individual data point [9]. The most typical approach is to add random perturbations, e.g., Gaussian [22], Laplacian [23], or Binomial noise [24], to the released statistics.

DP mechanisms have been investigated for FL under the assumption that the edge server is "honest-but-curious" and that communication is noiseless and unconstrained. In [25], Gaussian noise is added to the local model updates, and the power of the Gaussian noise is adapted to ensure a target privacy level. Analysis indicates that there is a tradeoff between convergence rate and privacy protection levels. Furthermore, a higher privacy guarantee is achievable if the DP algorithm uses random mini-batches — the so-called "privacy amplification by subsampling" principle [26]. Another DP mechanism based on random quantization is explored in [24], [27].

While the work reviewed so far assumes ideal communication, several recent works have appeared that share the common theme of exploiting the channel noise for differentially private FL. In [28], each device adds Gaussian noise before transmission via NOMA with static power allocation. The superposition property of NOMA is shown not only to provide benefits in terms of efficient gradient aggregation, but also to offer better privacy guarantees. Instead of injecting noise before transmission, an energy efficient approach is to scale down the transmit power [29]. A digital counterpart of these ideas is proposed in [30] which uses quantized gradient descent with privacy-inducing binomial noise. The quantization bits and noise parameters are optimized to maximize the convergence rate under channel capacity and privacy constraints.

All the discussed works assume a simple static power allocation, not accounting for the fact that channel noise has a different impact on convergence and privacy level. As our analysis demonstrates, channel noise added in the first iterations tends to impact convergence less significantly than the noise added in later iterations, whereas the privacy level depends on a weighted sum of the inverse noise power across the iteration. These properties, captured by compact analytical expressions derived in this paper, are leveraged to define optimization problems that are solved in closed form, yielding significant performance gains over standard static power allocation.

C. Contributions and Organization

In this paper, we study differentially private wireless decentralized gradient descent via the direct, uncoded, transmission of gradients from devices to edge server. The channel noise is utilized as a privacy preserving mechanism and dynamic power control is separately optimized for OMA and NOMA

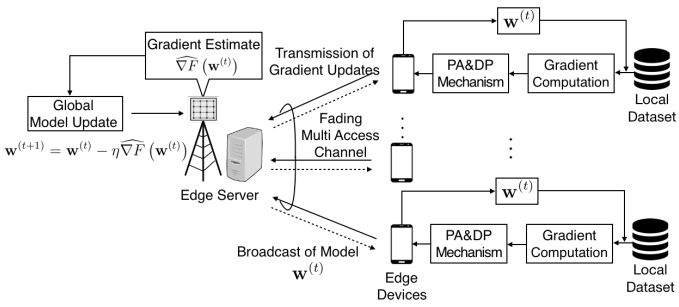


Figure 1. Differentially private federated edge learning system based on distributed gradient descent.

protocols with the goal of minimizing the learning optimality gap under privacy and power constraints across a given number of communication blocks. The main findings and contributions of the paper can be summarized as follows.

• **Offline optimized power allocation for OMA and NOMA:** Considering OMA and NOMA separately, we first analyze the convergence rate and privacy requirements for a given number of iterations under uncoded transmission. The resulting offline optimization problems are shown to be convex programs, and the optimal dynamic power allocation (PA) is obtained in closed form. The optimal PA is shown to be adaptive across the iterations, outperforming static PA assumed in prior works. The analytical results prove that privacy can be obtained "for free" as long as the privacy constraint level is below a threshold that decreases with the signal-to-noise (SNR). We also demonstrate that it is generally suboptimal to devote part of the transmitted power to actively add noise to the local updates. This is unlike the standard scenario with ideal communication, in which adding noise is essential to ensure DP constraints.

• **Online power allocation scheme:** A heuristic online approach is then proposed that leverages iterative one-step-ahead optimization based on the offline closed-form solutions, predicted channel state information (CSI).

• **Experiments:** We provides extensive numerical results that demonstrate the advantages of NOMA over conventional OMA protocols under DP constraints. We note here that these benefits is not a prior evident, since, with NOMA, devices transmit more frequently, and hence may leak more information if power is not properly allocated.

The remainder of the paper is organized as follows. Section II introduces the models and definitions. Section III presents the power allocation design for OMA. The design for NOMA is presented in Section IV. Section V provides numerical results, followed by conclusions in Section VI.

II. Models and Definitions

As shown in Fig. 1, we consider a wireless federated edge learning system comprising a single edge server and K edge devices connected through it via a shared noisy channel. Each device k has its own local dataset $\mathcal{D}_k$. This consists of labelled data samples $\{(\mathbf{u}_i, v_i)\} \in \mathcal{D}_k$, where $\mathbf{u}_i$ denotes the vector of covariates and v_i its associated label, which may be continuous or discrete. Local data sets are disjoint. A common regression or classification model, parameterized

by vector $\mathbf{w}$, is collaboratively trained by the edge devices through communications via edge server. In this section, we first introduce the learning protocol and the communication model, and then detail which the definition of differential privacy adopted in this work, along with main assumptions.

A. Learning Protocol

The regularized local loss function for the k-th device evaluated at model vector $\mathbf{w} \in \mathbb{R}^d$ is given by

$$\text{(Local loss function)} \quad F_k(\mathbf{w}) = \frac{1}{D_k} \sum_{(\mathbf{u},v)\in\mathcal{D}_k} f(\mathbf{w};\mathbf{u},v) + \lambda R(\mathbf{w}), \tag{1}$$

where $f(\mathbf{w};\mathbf{u},v)$ is the sample-wise loss function quantifying the prediction error of the model $\mathbf{w}$ on the training sample $\mathbf{u}$ with respect to (w.r.t.) its ground-truth label v; $D_k = |\mathcal{D}_k|$ is the cardinality of data set $\mathcal{D}_k$; and $R(\mathbf{w})$ is a strongly convex regularization function, which is scaled by hyperparameter $\lambda \geq 0$. The global loss function evaluated at model vector $\mathbf{w}$ is

$$\text{(Global loss function)} \quad F(\mathbf{w}) = \frac{1}{D_{\mathsf{tot}}} \sum_{k=1}^{K} D_k F_k(\mathbf{w}), \tag{2}$$

where $D_{\mathsf{tot}} = \sum_k D_k$. This amounts to the regularized empirical average of the sample-wise loss functions on the global data set $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$ obtained as the union of the local data sets. We note that the training loss (2) is an unbiased estimate of the generalization loss only if the devices observe independent and identically distributed (i.i.d.) samples from a common distribution. Nevertheless, this objective is also routinely considered for non-i.i.d. data sets in federated learning [31]–[33]. Other criteria that may be better suited to account for heterogeneous statistics across the devices may also be considered [34], but we leave this aspect for future work. The learning process aims to minimize the regularized global loss function as

$$\mathbf{w}^* = \arg\min F(\mathbf{w}). \tag{3}$$

In order to address problem (3), we study a differentially private implementation of federated distributed gradient descent via gradient-averaging. As we will detail, privacy is defined here from the point of view of any device with respect to the edge server, which is assumed to be "honest-but-curious". Accordingly, the edge server follows the protocol described below, but may attempt to infer information about data at the edge devices. We do not directly enforce privacy constraints on the other devices, which are implicitly trusted. More discussion on this point can be found in Section VI.

As illustrated in Fig. 1, at each t-th communication round, with $t = 1, \ldots, T$, the edge server broadcasts the current model iterate $\mathbf{w}^{(t)}$ to all edge devices via the downlink channel. We assume that downlink communication is ideal, so that each device receives the current model $\mathbf{w}^{(t)}$ without distortion. This assumption is practically well justified when the edge sever communicates through a base station with less stringent power constraint than the devices. By using the received current model $\mathbf{w}^{(t)}$ and the local dataset $\mathcal{D}_k$, each device computes the gradient of the local loss function in (1), that is

$$\text{(Local gradient)} \ \nabla F_k\big(\mathbf{w}^{(t)}\big) = \frac{1}{D_k} \sum_{(\mathbf{u},v)\in\mathcal{D}_k} \nabla f\big(\mathbf{w}^{(t)};\mathbf{u},v\big) + \lambda \nabla R(\mathbf{w}^{(t)}). \tag{4}$$

The devices transmit information about the local gradient (4) over the wireless shared channel to the edge server. Based on the received signal, the edge server obtains an estimate $\widehat{\nabla F}\left(\mathbf{w}^{(t)}\right)$ of the global gradient

$$\text{(Global gradient)} \quad \nabla F\left(\mathbf{w}^{(t)}\right) = \frac{1}{D_{\mathsf{tot}}} \sum_{k=1}^{K} D_k \nabla F_k\left(\mathbf{w}^{(t)}\right). \tag{5}$$

The edge server then updates the current global model via gradient descent

$$\text{(Model updating)} \quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \widehat{\nabla F}\left(\mathbf{w}^{(t)}\right), \tag{6}$$

where η denotes the learning rate. The steps in (4), (5), and (6) are iterated until a convergence condition is met.

The transmission of the gradient (4) from each k-th device may reveal information about the local data sets to the edge server. This motivates the use of DP as a rigorous mathematical framework to provide privacy guarantees that are agnostic to the computing resources and data processing requirements of the edge server. This will be detailed in Sec. II-C.

B. Communication Model

All devices communicate via the uplink to the edge server on the shared wireless channel using uncoded transmission. The main focus of this paper is the study of uncoded non-orthogonal multiple access (NOMA)[1] protocol, which enables over-the-air computing. For reference, we also study orthogonal multiple access (OMA) scheme under the same assumption of uncoded transmission. We note that it would be useful to include digital coded strategies for OMA as a benchmark. However, the design of digital communication protocols under DP constraints is a non-trivial problem that is currently subject to research [30].

We assume a block flat-fading channel, where the channel coefficients remain constant within a communication block, and they vary in a potentially correlated way over successive blocks. Each block contains d channel uses, allowing the uncoded transmission of a gradient vector. Due to memory and processing complexity constraints, on-device machine learning models are typically of small size, so that the model parameters dimension d can be assumed to be limited to a few tens of thousands of entries [35]. In this case, considering that typical coherence blocks may be of the same order of magnitude [36], [37], it is generally feasible to communicate the entire gradient vectors within one communication block. For larger model sizes, the gradient would need to be communicated across multiple coherence blocks – a setting that we leave

[1]In this context, NOMA is used as a transmission strategy, and it does not imply the use of specific decoders, such as successive decoding.

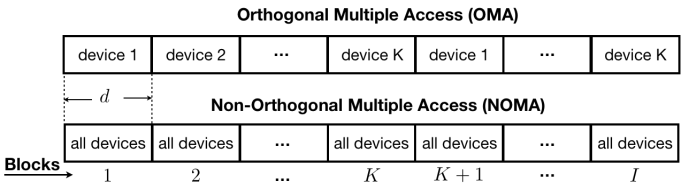


Figure 2. Illustration of the transmission schedule for the considered multiple access protocols.

for future investigations. We consider total I blocks available for training.

As in most papers on over-the-air computing [11]–[13], [15], [28], we assume perfect channel state information (CSI) at all nodes, so that each device can compensate for the phase of its own channel, ensuring the effective channel $h_k^{[i]}$ for each device k and block i is *real and non-negative*. This allows us to focus on a real channel model with non-negative channel gain, and it simplifies the design in power control parameters. We note that this assumption is also made in [14], [38], and that, as in these prior works, we do not make optimality claims in this regard. Details for OMA and NOMA are provided next.

1) Orthogonal Multiple Access (OMA): For orthogonal access, all devices time-share the channel uses via Time Division Multiple Access (TDMA). As illustrated in Fig. 2, devices are scheduled successively in orthogonal blocks, and we assume that the total number of blocks satisfies $I = KT$, so that T global gradient descent iterations (6) are implemented. In the i-th block, with $i = K(t-1)+k$, device k transmits gradient information corresponding to the t-th iteration. The signal received at the edge server during the i-th block is

$$\mathbf{y}_k^{[i]} = h_k^{[i]}\mathbf{x}_k^{[i]} + \mathbf{z}_k^{[i]}, \tag{7}$$

where $h_k^{[i]} \geq 0$ is the channel gain for device k in block i, $\mathbf{x}_k^{[i]} \in \mathbb{R}^d$ is an uncoded function of the local gradient $\nabla F_k(\mathbf{w}^{(t)})$, and $\mathbf{z}_k^{[i]}$ is channel noise i.i.d. according to distribution $\mathcal{N}(0, N_0\mathbf{I})$. We define as $\mathbf{y}^{(t)} = [\mathbf{y}^{[K(t-1)+1]}, \cdots, \mathbf{y}^{[Kt]}]$ the vector collecting all signals received for iteration t across K blocks.

2) Non-Orthogonal Access: For non-orthogonal access, we assume symbol-level synchronization among the devices that transmit simultaneously in each block. This can be achieved by using standard protocols such as the timing advance procedure in LTE and 5G NR [39]. In the i-th block, all devices upload the local gradients corresponding to the $t = i$-th iteration, and we have $I = T$ so that the number of blocks equals the number of iterations. The corresponding received signal is

$$\mathbf{y}^{[i]} = \sum_{k=1}^{K} h_k^{[i]}\mathbf{x}_k^{[i]} + \mathbf{z}^{[i]}, \tag{8}$$

where $h_k^{[i]}$ and $\mathbf{z}^{[i]}$ are defined as above; and signal $\mathbf{x}_k^{[i]} \in \mathbb{R}^d$ encodes information about the local gradient $\nabla F_k(\mathbf{w}^{(t)})$ with $t = i$. For NOMA, we will also write $\mathbf{y}^{(t)} = \mathbf{y}^{[t]}$.

Note that for both forms of access, the transmit power constraint of a device is given as

$$\text{(Power constraint)} \quad \mathsf{E}[\|\mathbf{x}_k^{[i]}\|^2] \leq P. \tag{9}$$

Accordingly, we define the maximum signal to noise ratio (SNR) as

$$\mathsf{SNR}_{\mathsf{max}} = \frac{P}{dN_0}, \tag{10}$$

where dN_0 represents the power of the channel nosies within one communication block. We refer to (10) as the maximum SNR since devices may optimally transmit with a power strictly smaller than P in (9) in order to satisfy the DP constraints.

C. Differential Privacy

As a threat model, we assume a "honest-but-curious" edge server that may attempt to infer information about local data sets from the signals $\{\mathbf{y}^{(t)}\}_{t=1}^T$ received across T successive iterations. Note that, as discussed, T iterations correspond to T communication blocks for NOMA and TK blocks for OMA. The standard definition DP imposes a point-wise upper bound on the divergence between the distributions $P(\mathbf{y}|\mathcal{D})$ and $P(\mathbf{y}|\mathcal{D}')$ of the received signals $\mathbf{y} = \{\mathbf{y}^{(t)}\}_{t=1}^T$ conditioned on the use of either one of two "neighboring" global data sets $\mathcal{D}$ and $\mathcal{D}'$. The two neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$ differ only by one sample at one of the devices. Defining the cardinality of the set difference for two sets $\mathcal{A}$ and $\mathcal{B}$ as $\|\mathcal{A} - \mathcal{B}\|_1$, we have the following formal definition.

Definition 1 (Differential Privacy [9])**.** The communication and learning protocol is (ϵ, δ)-differentially private, where $\epsilon > 0$, and $\delta \in [0, 1)$, if any two possible adjacent global datasets $\mathcal{D}' = \bigcup_{k=1}^K \mathcal{D}_k'$ and $\mathcal{D}'' = \bigcup_{k=1}^K \mathcal{D}_k''$, with $\left\|\mathcal{D}_j' - \mathcal{D}_j''\right\|_1 = 1$ for some device j and $\left\|\mathcal{D}_k' - \mathcal{D}_k''\right\|_1 = 0$ for all $k \neq j$, we have the inequality

$$P(\mathbf{y}|\mathcal{D}') \leq \exp(\epsilon)P(\mathbf{y}|\mathcal{D}'') + \delta. \tag{11}$$

The bound (11) can be interpreted in terms of the test variable.

$$\text{(Differential privacy loss)} \quad \mathcal{L}_{\mathcal{D}',\mathcal{D}''}(\mathbf{y}) = \ln \frac{P(\mathbf{y}|\mathcal{D})}{P(\mathbf{y}|\mathcal{D}')}, \tag{12}$$

which is referred to as differential privacy loss. This corresponds to the log-likelihood ratio for the detection of neighboring data sets $\mathcal{D}'$ and $\mathcal{D}''$. The (ϵ, δ)-DP condition (11) ensures that, for all possible adjacent global datasets $\mathcal{D}'$ and $\mathcal{D}''$, the absolute value of privacy loss variable (12) is bounded by ϵ with probability at least $1 - \delta$, i.e., $\Pr(|\mathcal{L}_{\mathcal{D}',\mathcal{D}''}(\mathbf{y})| \leq \epsilon) \geq 1 - \delta$ (see Lemma 3.17 in [9]). If ϵ and δ are suitably small, this makes it statistically impossible, even for an adversary that knows all data points in $\mathcal{D}$ except one, to identify the remaining individual sample.

D. Assumptions On the Loss Functions

Finally, we list several standard assumptions we make on the loss functions and on its gradients.

Assumption 1 (Smoothness)**.** The global loss function $F(\mathbf{w})$ is smooth with constant $L > 0$, that is, it is continuously differentiable and the gradient $\nabla F(\mathbf{w})$ is Lipschitz continuous with constant L, i.e.,

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \text{for all } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{13}$$

Inequality (13) implies the following inequality

$$F(\mathbf{w}') \leq F(\mathbf{w}) + \nabla F(\mathbf{w})^{\mathsf{T}}(\mathbf{w}' - \mathbf{w}) + \frac{L}{2}\left\|\mathbf{w} - \mathbf{w}'\right\|^2, \quad \text{for all } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{14}$$

Assumption 2 (Polyak-Lojasiewicz Inequality)**.** The optimization problem (3) has a non-empty solution set. Furthermore, denoting as F^* the corresponding optimal function value, the global loss function $F(\mathbf{w})$ satisfies the Polyak-Lojasiewicz (PL) condition, that is, the following inequality holds for some constant $\mu > 0$

$$\frac{1}{2}\left\|\nabla F(\mathbf{w})\right\|^2 \geq \mu\left[F(\mathbf{w}) - F^*\right]. \tag{15}$$

The PL condition (15) is significantly more general than the standard assumption of strong convexity [40]. A strong convex with constant $\mu > 0$ implies the PL inequality with same parameter μ [41]. Note also that for a convex loss function $f(\cdot; \mathbf{u}, v)$, e.g., for least squares and logistic regression, the strong convexity of function $F(\mathbf{w})$ follows from the addition of the regularizing term with $\lambda > 0$.

III. Orthogonal Multiple Access

In this section, we consider the design and analysis of orthogonal multiple access with uncoded transmission and adaptive power control. To start, we assume that, at each iteration t, device k transmits a scaled and noisy version of the gradient $\mathbf{x}_k^{[i]} = \mathbf{x}_k^{(t)}$ in block $i = K(t-1) + k$ as

$$\mathbf{x}_k^{(t)} = \alpha_k^{(t)}\left(D_k \nabla F_k\left(\mathbf{w}^{(t)}\right) + \mathbf{n}_k^{(t)}\right). \tag{16}$$

In (16), the artificial noise term $\mathbf{n}_k^{(t)} \sim \mathcal{N}(0, (\sigma_k^{(t)})^2\mathbf{I})$ is added in accordance to the standard Gaussian mechanism in the DP literature; and $\alpha_k^{(t)} \geq 0$ is a scaling factor. We note that, by (16), the effective noise in the received signal (7) is given by the summation of channel and artificial noise. The standard deviation of the effective noise is

$$m_k^{(t)} = \sqrt{(h_k^{(t)}\alpha_k^{(t)}\sigma_k^{(t)})^2 + N_0}. \tag{17}$$

We are interested in optimizing over the sequences of parameters $(\alpha_k^{(1)},\cdots,\alpha_k^{(T)})$ and $(\sigma_k^{(1)},\cdots,\sigma_k^{(T)})$ in order to maximize the learning performance under the (ϵ, δ)-DP constraint. To this end, the remainder of this section first provides DP and convergence analysis, based on which the optimization problem is then formulated and solved. Throughout this section, we use the notation $h_k^{[i]} = h_k^{(t)}$, $\mathbf{y}_k^{[i]} = \mathbf{y}_k^{(t)}$ and $\mathbf{z}_k^{[i]} = \mathbf{z}_k^{(t)}$ for $i = K(t-1) + k$, and we make the following common assumption (see, e.g., [23], [42], [43]).

Assumption 3 (Bounded Sample-Wise Gradient)**.** At any iteration t, for any training sample $(\mathbf{u}, v)$, the gradient is upper bounded by a given constant $\gamma^{(t)}$, i.e., for all possible $(\mathbf{u}, v)$ (not limited to those in data sets $\{\mathcal{D}_k\}$) we have the inequality

$$\left\|\nabla f(\mathbf{w}^{(t)}; \mathbf{u}, v)\right\| \leq \gamma^{(t)}. \tag{18}$$

A. Differential Privacy Analysis

By standard results on DP, the privacy level (ϵ, δ) depends on the sensitivity of the function being disclosed, excluding the effect of noise, to the input data set. More specifically, the sensitivity measures the amount by which a single individual data point can change the disclosed function in the worst case. For each device k, the edge server is assumed to be informed about parameters $\{\alpha_k^{(t)}\}$. We assume here that those parameters are fixed constants that do not reveal information about the local datasets. Hence, the only function of the data being disclosed is the received signal $\mathbf{y}_k^{(t)}$, upon subtraction of the effective noise. The sensitivity $\Delta_k^{(t)}$ of the noiseless received signal $\mathbf{y}_k^{(t)} - \mathbf{z}_k^{(t)} - h_k^{(t)}\alpha_k^{(t)}\mathbf{n}_k^{(t)}$ is defined as

$$\text{(Sensitivity in OMA) } \Delta_k^{(t)} = \max_{\mathcal{D}_k', \mathcal{D}_k''} \left\| h_k^{(t)}\alpha_k^{(t)} \times \left(\sum_{(\mathbf{u},v)\in\mathcal{D}_k'} \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big) - \sum_{(\mathbf{u},v)\in\mathcal{D}_k''} \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big)\right)\right\|, \tag{19}$$

where data sets $\mathcal{D}_k'$ and $\mathcal{D}_k''$ satisfy $\|\mathcal{D}_k' - \mathcal{D}_k''\|_1 = 1$. By the triangular inequality and Assumption 3, we have the bound

$$\Delta_k^{(t)} \leq 2h_k^{(t)}\alpha_k^{(t)}\gamma^{(t)}. \tag{20}$$

Lemma 1 (Differential Privacy Guarantees for OMA)**.** For any fixed sequence of parameters $\{\alpha_k^{(t)}, \sigma_k^{(t)}\}$, federated gradient averaging via OMA guarantees (ϵ, δ)-DP if the following condition is satisfied

$$\sum_{t=1}^{T}\left(\frac{\sqrt{2}h_k^{(t)}\alpha_k^{(t)}\gamma^{(t)}}{m_k^{(t)}}\right)^2 \leq \left(\sqrt{\epsilon + \left[\mathcal{C}^{-1}\left(1/\delta\right)\right]^2} - \mathcal{C}^{-1}\left(1/\delta\right)\right)^2 \tag{21}$$

$$\triangleq \mathcal{R}_{\mathsf{dp}}(\epsilon, \delta), \text{ for all } k, \tag{22}$$

where $m_k^{(t)}$ in (17) is the standard deviation of the effective noise, and $\mathcal{C}^{-1}(x)$ is the inverse function of $\mathcal{C}(x) = \sqrt{\pi}xe^{x^2}$.

Proof: The proof is based on the advanced composition theorem [9, Theorem 3.20] and is detailed in Appendix A.

Lemma 1, along with (20), indicate that the privacy level depends on the sum of the per-iteration ratios $(\sqrt{2}h_k^{(t)}\alpha_k^{(t)}\gamma^{(t)}/m_k^{(t)})^2$, which, by (16), depend on the ratio between useful signal and effective noise powers. The effective noise level $m_k^{(t)}$ contributing to the privacy of device k equals the sum of the channel noise power and of the noise added by device k in (16). The constraint (21) suggests that the effective noise variance can be adapted to the sequence of channel gains, as long as the impact on convergence is suitably accounted for.

B. Convergence Analysis

At the t-th iteration, encompassing the blocks $i = K(t-1)+1, \cdots, Kt$, the edge server estimates the scaled local gra-

dient $D_k \nabla F_k(\mathbf{w}^{(t)})$ as $(h_k^{(t)}\alpha_k^{(t)})^{-1}\mathbf{y}_k^{(t)}$, and then the global gradient is estimated as

$$\begin{aligned}\widehat{\nabla F}(\mathbf{w}^{(t)}) &= \frac{1}{D_{\mathsf{tot}}}\sum_{k=1}^{K}\left(h_k^{(t)}\alpha_k^{(t)}\right)^{-1}\mathbf{y}_k^{(t)} \\ &= \frac{1}{D_{\mathsf{tot}}}\sum_{k=1}^{K} D_k \nabla F_k\left(\mathbf{w}^{(t)}\right) + \mathbf{n}_k^{(t)} + \left(h_k^{(t)}\alpha_k^{(t)}\right)^{-1}\mathbf{z}_k^{(t)}. \end{aligned} \tag{23}$$

Building on standard results on gradient descent with noisy gradient [41], we have the following bound on the average optimality gap at the end of iteration T.

Lemma 2 (Optimality Gap Bound for OMA)**.** Under Assumptions 1 and 2, for a learning rate $\eta = 1/L$, after T iterations the average optimality gap is upper bounded as

$$\begin{aligned}\mathsf{E}\left[F\left(\mathbf{w}^{(T+1)}\right) - F^*\right] \leq \left(1-\frac{\mu}{L}\right)^{T}\left[F\left(\mathbf{w}^{(1)}\right) - F^*\right] \\ + \frac{d}{2LD_{\mathsf{tot}}^2}\sum_{t=1}^{T}\left(1-\frac{\mu}{L}\right)^{T-t}\sum_{k=1}^{K}\left(\frac{m_k^{(t)}}{h_k^{(t)}\alpha_k^{(t)}}\right)^2,\end{aligned} \tag{24}$$

where the standard deviation $m_k^{(t)}$ of the effective noise is defined in (17).

Proof: See Appendix B.

The first term in (24) reflects the standard geometric decay of the initial optimality gap $(F(\mathbf{w}^{(1)}) - F^*)$ as T increases, while the second accounts for the impact of the effective additive noise powers (21). Interestingly, the bound (24) suggests that noise added in the initial iterations is less damaging to the final optimality gap than the noise added in later iterations. This is because the contribution of the noise added at iteration t is discounted by a factor $(1-\mu/L)^{T-t}$. We will leverage this result in the next section to optimize power allocation.

C. Optimization

In this section, we are interested in minimizing the optimality bound in Lemma 2 under (ϵ,δ)-DP constraint (21) and the power constraints (9), for all K devices across T iterations. Note that, for the objective function (24), the optimization variables only exist in the second term. By replacing $m_k^{(t)}$ with its definition given in (17), the resulting optimization problem (**OMA Opt.**) of interest is formulated as

$$\min_{\{\sigma_k^{(t)},\alpha_k^{(t)}\}_{k=1}^{K}} \sum_{t=1}^{T}\left(1-\frac{\mu}{L}\right)^{-t}\sum_{k=1}^{K}\left[(\sigma_k^{(t)})^2 + \left(\frac{\sqrt{N_0}}{h_k^{(t)}\alpha_k^{(t)}}\right)^2\right] \tag{25a}$$

$$\text{s.t.} \sum_{t=1}^{T}\frac{(\sqrt{2}\gamma^{(t)})^2}{(\sigma_k^{(t)})^2 + N_0/(h_k^{(t)}\alpha_k^{(t)})^2} \leq \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta),\ \forall k, \tag{25b}$$

$$(\alpha_k^{(t)})^2\left[(D_k G_k^{(t)})^2 + d(\sigma_k^{(t)})^2\right] \leq P, \quad \forall k,t, \tag{25c}$$

where $\mathcal{R}_{\mathsf{dp}}(\epsilon,\delta)$ is defined in (22), and Parameter $G_k^{(t)}$ represents an upper bound on the norm of the local gradient as $\left\|\nabla F_k\left(\mathbf{w}^{(t)}\right)\right\| \leq G_k^{(t)}$. By Assumption 3, we have $G_k^{(t)} \leq \gamma^{(t)}$. Under OMA, the optimization (25a)-(25c) over the additive noise deviations $\{\sigma_k^{(1)},\cdots,\sigma_k^{(T)}\}$ and scaling factors $\{\alpha_k^{(1)},\cdots,\alpha_k^{(T)}\}$ for each devices k can be carried out in parallel. The corresponding problem (**OMA Local Opt.**) to be solved by device k is

$$\min_{\{\sigma_k^{(t)},\alpha_k^{(t)}\}} \sum_{t=1}^{T}\left(1-\frac{\mu}{L}\right)^{-t}\left[(\sigma_k^{(t)})^2 + \left(\frac{\sqrt{N_0}}{h_k^{(t)}\alpha_k^{(t)}}\right)^2\right] \tag{26a}$$

$$\text{s.t.} \sum_{t=1}^{T}\frac{(\sqrt{2}\gamma^{(t)})^2}{(\sigma_k^{(t)})^2 + N_0/(h_k^{(t)}\alpha_k^{(t)})^2} \leq \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta), \tag{26b}$$

$$(\alpha_k^{(t)})^2\left[(D_k G_k^{(t)})^2 + d(\sigma_k^{(t)})^2\right] \leq P, \quad \forall t. \tag{26c}$$

Without the DP constraint (26b), the optimal solution to problem (26) is to fully use the power budget P for the transmission of the local gradient, i.e., to set $\alpha_k^{(t)} = \sqrt{P}/(D_k G_k^{(t)})$ and $\sigma_k^{(t)} = 0$ for all k and t. Due to the DP constraint, we now show that this may not be the optimal solution if the privacy condition is sufficiently strict.

Before detailing offline and online solutions, it is useful to observe that, in order for constraint (26b) to guarantee (ϵ,δ)-DP, by leave t, it is necessary that the parameters $G_k^{(t)}$ be fixed at each iteration t in a way that does not depend on the local data sets. We will return to this point when discussing online methods.

1) Offline Optimization: We first assume that the parameters $\{h_k^{(t)},\gamma^{(t)},G_k^{(t)}\}$ are known beforehand so that problem (26a)-(26c) can be tackled offline. As we show in Appendix C, problem (26a)-(26c) can be converted into a convex program via a change of variables. The resulting optimal solution is described in the following theorem.

Theorem 1. The optimal offline solution of problem (25a)-(25c) under OMA is given as follows:

- If condition

$$\sum_{t=1}^{T}\frac{P(\sqrt{2}\gamma^{(t)}h_k^{(t)})^2}{N_0(D_k G_k^{(t)})^2} < \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta) \tag{27}$$

  holds, there exists a unique optimal solution given as $(\alpha_k^{(t)})_{\mathsf{opt}} = \sqrt{P}/(D_k G_k^{(t)})$ and $(\sigma_k^{(t)})_{\mathsf{opt}} = 0$. In this case, the power budget P is fully used for the transmission of the local gradient, and the channel noise is sufficient to guarantee privacy. The optimal solution is identical as that of without DP constraint, and privacy is hence obtained "for free";
- Otherwise, there exist multiple optimal solutions, and the solution that minimizes the transmit power is

$$(\alpha_k^{(t)})_{\mathsf{opt}} = \min\left\{\frac{\sqrt{N_0}(2\zeta_k)^{-\frac{1}{4}}}{h_k^{(t)}\sqrt{\gamma^{(t)}}}\left(1-\frac{\mu}{L}\right)^{-t/4}, \frac{\sqrt{P}}{D_k G_k^{(t)}}\right\} \tag{28}$$

$$(\sigma_k^{(t)})_{\mathsf{opt}} = 0, \tag{29}$$

  where the value of parameter ζ_k can be obtained by bisection to satisfy the constraint

$$\sum_{t=1}^{T}(\sqrt{2}\gamma^{(t)})^2 \min\left\{\frac{(1-\mu/L)^{-t/2}}{\sqrt{2\zeta_k}\gamma^{(t)}}, \frac{P(h_k^{(t)})^2}{(\sqrt{N_0}D_k G_k^{(t)})^2}\right\} = \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta). \tag{30}$$

In this case, the transmitted power needs to be scaled down in order to leverage the channel noise to ensure (ϵ, δ)-DP.

Proof: The proof is detailed in Appendix C.

A first interesting observation from Theorem 1 is that it is optimal for the devices not to add noise to the transmitted signals (16). It is, in fact, sufficient to scale down their transmitted powers, via the choice of $\alpha_k^{(t)}$, with smaller powers transmitted when more stringent DP constraints are imposed. Second, when condition (27) is satisfied, privacy is obtained "for free", that is, without affecting the learning performance of the system, as the devices can use their full power. Third, condition (27) is less strict as D_k increases, showing that devices with larger datasets can attain privacy "for free" over a broader range of SNR levels. Finally, we note that it is generally suboptimal to use a time-invariant policy that sets the scaling factor $\alpha_k^{(t)}$ as a constant.

2) Online Optimization: Theorem 1 assumes that the sequence of parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$ is known a priori so as to enable offline optimization. Here we describe a heuristic online approach that leverages iterative one-step-ahead optimization based on predicted values for the future parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$.

To elaborate, assume that, at each iteration t, we have predicted values $\{\widehat{h}_k^{(t')}, \widehat{\gamma}^{(t')}, \widehat{G}_k^{(t')}\}$ for $t' = t, t+1, \cdots, T$ and that the accumulated DP loss is given by $\mathcal{L}_k^{(t-1)} = \sum_{t'=1}^{t-1} (\sqrt{2} h_k^{(t')} \alpha_k^{(t')} \gamma^{(t')} / m_k^{(t')})^2$ from (21). As summarized in Algorithm 1, we propose to apply the solution in Theorem 1 to the interval $(t, t+1, \cdots, T)$ by replacing the true parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$ with the estimates $\{\widehat{h}_k^{(t)}, \widehat{\gamma}^{(t)}, \widehat{G}_k^{(t)}\}$ and the DP constraint with the residual $\mathcal{R}_{\mathsf{dp}}(\epsilon, \delta) - \mathcal{L}_k^{(t-1)}$. The produced scaling factors $\alpha_k^{(t)}$ are then applied, and the procedure is repeated for iteration $t+1$. We now discuss the problem of prediction of parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$.

To start, we model the sequence of fading channels $\{g_k^{[i]}\}$ via an autoregressive (AR) Rician model. We note that the method can be directly extended to other probabilistic models. Accordingly, each channel gain $h_k^{[i]}$ is obtained as $h_k^{[i]} = |g_k^{[i]}|$, where the complex channel coefficient $\{g_k^{[i]}\}$ is given as

$$g_k^{[i]} = \sqrt{\frac{\kappa_k}{\kappa_k + 1}} + \sqrt{\frac{1}{\kappa_k + 1}} r_k^{[i]}, \tag{31}$$

with κ_k being the Rice parameter and the stochastic diffuse component $r_k^{[i]}$ following an $AR(1)$ process. We specifically write $r_k^{[i+1]} = \rho_k r_k^{[i]} + \sqrt{1 - \rho_k^2} \widetilde{r}_k^{[i]}$, with temporal correlation coefficient $0 \leq \rho_k \leq 1$, and $\widetilde{r}_k^{[i]} \sim \mathcal{CN}(0, 1)$ being an i.i.d. innovation process. Given the current CSI $g_k^{[i]}$, the future channel power $(h_k^{[j]})^2 = |g_k^{[i]}|^2$ for $j > i$ can be predicted via minimum mean squared error (MMSE) estimation as

$$(\widehat{h}_k^{[j]})^2 = \mathsf{E}\big[(h_k^{[j]})^2 \big| g_k^{[i]}\big] = \frac{\kappa_k + (\rho_k^{j-i})^2}{\kappa_k + 1} |g_k^{[i]}|^2 + \frac{1 - (\rho_k^{j-i})^2}{\kappa_k + 1}. \tag{32}$$

Next, we discuss the estimations of parameters $\{\gamma^{(t)}\}$ and $\{G_k^{(t)}\}$. Parameter $\gamma^{(t)}$ is by definition independent of the local data sets, and is typically determined by clipping the local gradient before transmission to the server [22], [44]. To this end, in (4), we substitute the per-sample gradient $\nabla f\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right)$ with its clipped version

(Clipped per-sample gradient)

$$\overline{\nabla f}\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right) = \min\left\{1, \frac{\widehat{\gamma}}{\|\nabla f\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right)\|}\right\} \times \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big) \tag{33}$$

for some fixed threshold $\widehat{\gamma} > 0$.

The definition of the parameters $\{G_k^{(t)}\}$ makes them generally data-dependent. In order to avoid leaking additional information about the data to the server, we propose to predict bounds $\{\widehat{G}_k^{(t')}\}$ for $t' \geq t$ based on an additional signal broadcast by the server. Specifically, we let the edge server transmit the positive scalar $\|\mathbf{y}_k^{(t-1)}\|/(h_k^{(t-1)} \alpha_k^{(t-1)})$ back to device k in addition to the broadcast signal $\mathbf{w}^{(t)}$. Basing the predictions $\widehat{G}_k^{(t')}$ on the past received signal $\mathbf{y}_k^{(t-1)}$ does not affect privacy, since the privacy loss due to the reception of $\mathbf{y}_k^{(t-1)}$ at the edge server is accounted for by $\mathcal{L}_k^{(t-1)}$. At each iteration t, any device k sets

$$\widehat{G}_k^{(t)} = \begin{cases} \|\mathbf{y}_k^{(t-1)}\|/(h_k^{(t-1)} \alpha_k^{(t-1)} D_k), & t > 1, \\ \widehat{\gamma}, & t = 1. \end{cases} \tag{34a}$$

and $\widehat{G}_k^{(t')} = \widehat{G}_k^{(t)}, \forall\ t' > t.$ (34b)

Furthermore, in order to ensure constraint on the bounded local gradient $\left\|\nabla F_k\left(\mathbf{w}^{(t)}\right)\right\| \leq G_k^{(t)}$, we clip the local gradient for transmission as

(Clipped gradient transmission in OMA)

$$\overline{\nabla f}\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right) = \min\left\{1, \frac{\widehat{\gamma}}{\|\nabla f\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right)\|}\right\} \times \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big) \tag{35}$$

with $\overline{\nabla F}_k = \frac{1}{D_k} \sum_{(\mathbf{u}, v) \in \mathcal{D}_k} \overline{\nabla f}\left(\mathbf{w}^{(t)}; \mathbf{u}, v\right) + \lambda \nabla R(\mathbf{w})$.

Finally, we observe that, strictly speaking, the analysis of convergence in Lemma 2 should be modified in order to account for clipping, but we found the heuristic approach summarized in Algorithm 1 to perform well in practice.

IV. Non-orthogonal Multiple Access

In this section, we consider the design and analysis of NOMA. For the t-th iteration, local gradients are transmitted using the uncoded strategy (16). As in [11]–[13], [15], [28], we select the scaling factors $\alpha_k^{(t)}$ so as to ensure that, in the absence of noise, the edge server can recover a scaled version of the global gradient (5). Accordingly, we set

$$\text{(Gradient Alignment)} \quad h_k^{(t)} \alpha_k^{(t)} = c^{(t)}, \tag{36}$$

for some constant $c^{(t)}$. We note that the effective noise in NOMA is equal to the summation of the channel noise and the contributions of artificial noise from all devices, and its standard deviation is given by

$$m^{(t)} = \sqrt{(c^{(t)})^2 \sum_{k=1}^{K} (\sigma_k^{(t)})^2 + N_0}. \tag{37}$$

Algorithm 1 Online Scheme for OMA

Input: DP level $\mathcal{R}_{\mathsf{dp}}$, channel noise $\sqrt{N_0}$, channel correlation ρ, clipping threshold $\gamma^{(t)} = \widehat{\gamma}$
Initialize: Local privacy loss $\mathcal{L}_k^{(0)} = 0$
For each iteration: $t = 1, \ldots, T$
  For each device: $k = 1, \ldots, K$
    Receive $\mathbf{w}^{(t)}$ from edge server
    Update local model by (35)
    If $t > 1$
      Receive $\mathbf{y}_k^{(t-1)}/(h_k^{(t-1)}\alpha_k^{(t-1)})$ from edge server
    end
    Compute predictors $\{\widehat{h}_k^{(t')}, \widehat{G}_k^{(t')}\}$ via (32) and (34) for $t' \in [t, \cdots, T]$ with $\widehat{h}_k^{(t)} = h_k^{(t)}$
    Apply Theorem 1 over the time interval $[t, \cdots, T]$ with $\{h_k^{(t')} \leftarrow \widehat{h}_k^{(t')},\ G_k^{(t')} \leftarrow \widehat{G}_k^{(t')},\ \gamma^{(t)} \leftarrow \widehat{\gamma}\}$ and residual DP constraint $\mathcal{R}_{\mathsf{dp}} - \mathcal{L}_k^{(t-1)}$
    Use optimized scaling factor $\alpha_k^{(t)}$ to transmit (35)
    Update local privacy loss as $\mathcal{L}_k^{(t)} = \mathcal{L}_k^{(t-1)} + \frac{(\sqrt{2}\gamma^{(t)} h_k^{(t)} \alpha_k^{(t)})^2}{N_0}$
  end
end

As per (36), in this section, we are optimizing over the parameters $(c^{(1)}, \cdots, c^{(T)})$ as well as over the added noise power $(\sigma_k^{(1)}, \cdots, \sigma_k^{(T)})$. Throughout this section, we denote $h_k^{[i]} = h_k^{(t)}$, and $\mathbf{z}^{[i]} = \mathbf{z}^{(t)}$ for $i = t$.

A. Differential Privacy Analysis

As discussed in Section III-A, the DP level depends on the sensitivity of the function being disclosed, which, in NOMA, for the same reasons discussed in Section III-A, is the received noiseless aggregated signal. The sensitivity to change in the data set of device k is accordingly defined as

$$\text{(Sensitivity in NOMA)}\quad \Delta_k^{(t)} = \max_{\mathcal{D}_k', \mathcal{D}_k''} \Big\| c^{(t)} \times \Big(\sum_{(\mathbf{u},v)\in\mathcal{D}'} \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big) - \sum_{(\mathbf{u},v)\in\mathcal{D}''} \nabla f\big(\mathbf{w}^{(t)}; \mathbf{u}, v\big) \Big) \Big\|, \quad (38)$$

where $\|\mathcal{D}_k' - \mathcal{D}_k''\|_1 = 1$, $\left\|\mathcal{D}_j' - \mathcal{D}_j''\right\|_1 = 0$ for all $j \neq k$, and $\mathcal{D}' = \bigcup_{k=1}^K \mathcal{D}_k'$, $\mathcal{D}'' = \bigcup_{k=1}^K \mathcal{D}_k''$. By Assumption 3, we can bound the sensitivity as

$$\Delta_k^{(t)} \leq 2c^{(t)}\gamma^{(t)}. \quad (39)$$

Then, the DP guarantees for NOMA are given as follows.

Lemma 3 (Differential Privacy Guarantees for NOMA)**.** Federated gradient averaging via NOMA guarantees (ϵ, δ)-DP if the following condition is satisfied

$$\sum_{t=1}^T \left(\frac{\sqrt{2}c^{(t)}\gamma^{(t)}}{m^{(t)}} \right)^2 \leq \mathcal{R}_{\mathsf{dp}}(\epsilon, \delta), \text{ for all } k. \quad (40)$$

where $m^{(t)}$ is the standard deviation of the effective noise (37).
Proof: The proof follows in a manner similar to Lemma 1 by replacing the sensitivity and effective noise with those defined in NOMA.

Lemma 3 indicates that the effective noise contributing to the privacy of each device k is given by the sum of channel noise and the privacy-inducing noise added by all devices. This is an important advantage of NOMA, which was also observed in [28].

B. Convergence Analysis

At the t-th iteration, the edge server estimates the global gradient as

$$\widehat{\nabla F}(\mathbf{w}^{(t)}) = \frac{1}{D_{\mathsf{tot}}} \left(c^{(t)}\right)^{-1} \mathbf{y}^{(t)} = \frac{1}{D_{\mathsf{tot}}} \sum_{k=1}^K D_k \nabla F_k\big(\mathbf{w}^{(t)}\big) + \mathbf{n}_k^{(t)} + \left(c^{(t)}\right)^{-1} \mathbf{z}_k^{(t)}. \quad (41)$$

Lemma 4 (Optimality Gap Bound for NOMA)**.** Under Assumptions 1 and 2, for a learning rate $\eta = 1/L$, after T iterations the average optimality gap is upper bounded as

$$\mathsf{E}\left[F\left(\mathbf{w}^{(T+1)}\right) - F^*\right] \leq \left(1 - \frac{\mu}{L}\right)^T \left[F\big(\mathbf{w}^{(1)}\big) - F^*\right] + \frac{d}{2LD_{\mathsf{tot}}^2} \sum_{t=1}^T \left(1 - \frac{\mu}{L}\right)^{T-t} \left(\frac{m^{(t)}}{c^{(t)}}\right)^2, \quad (42)$$

where the standard deviation $m^{(t)}$ of the effective noise is defined in (37).
Proof: The proof follows via the same steps reported in Appendix B by replacing the summation of (17) with (37).

C. Optimization

In this section, we are interested in minimizing the optimality bound in Lemma 4 under the (ϵ, δ)-DP constraint (40) and the power constraints (9) across T iterations. The resulting optimization problem for NOMA (**NOMA Opt.**) is formulated as

$$\min_{\{\sigma_k^{(t)}, c^{(t)}\}_{k=1}^K} \sum_{t=1}^T \left(1 - \frac{\mu}{L}\right)^{-t} \sum_{k=1}^K (\sigma_k^{(t)})^2 + N_0/(c^{(t)})^2 \quad (43a)$$

$$\text{s.t.} \sum_{t=1}^T \frac{(\sqrt{2}\gamma^{(t)})^2}{\sum_{k=1}^K (\sigma_k^{(t)})^2 + N_0/(c^{(t)})^2} \leq \mathcal{R}_{\mathsf{dp}}(\epsilon, \delta) \quad (43b)$$

$$\left(\frac{c^{(t)}}{h_k^{(t)}}\right)^2 \left[(D_k G_k^{(t)})^2 + d(\sigma_k^{(t)})^2\right] \leq P, \ \forall k, t. \quad (43c)$$

Without the DP constraint (43b), the optimal solution to problem (43) is determined by the devices with the smallest value of the ratio $h_k^{(t)}/(D_k G_k^{(t)})$ due to the need to satisfy the gradient alignment condition (36). In particular, the optimal solution prescribes that such devices use the full power budget P to transmit the local gradient while the other devices transmit at the maximum power allowed under condition (36), i.e., $c^{(t)} = \sqrt{P}\min_k h_k^{(t)}/(D_k G_k^{(t)})$ and $\sigma_k^{(t)} = 0$. We will see next that this is no longer the optimal solution under sufficiently strict DP constraints.

1) Offline Optimization: We first assume that the parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$ are known beforehand so that problem (43a)-(43c) can be tackled offline. With a change of the variables, the problem can be shown to be convex. Unlike the previously studied problem for OMA, the optimization (43a)-(43c) cannot be solved in parallel across the devices.

Theorem 2. The optimal offline solution of problem (43a)-(43c) under NOMA is given as follows:

- If condition

$$\frac{2P}{N_0}\sum_{t=1}^{T}(\gamma^{(t)})^2 \min_k \left(\frac{h_k^{(t)}}{D_k G_k^{(t)}}\right)^2 < \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta) \quad (44)$$

  holds, there exists a unique optimal solution given as $(c^{(t)})_{\mathsf{opt}} = \sqrt{P}\min_k h_k^{(t)}/(D_k G_k^{(t)})$ and $(\sigma_k^{(t)})_{\mathsf{opt}} = 0$. In this case, the devices with smallest value of the ratio $h_k^{(t)}/(D_k G_k^{(t)})$ transmit using the full power budget P, while the other devices do not use full power. Therefore, under the gradient alignment condition (36), privacy is obtained "for free";
- Otherwise, there exist multiple optimal solutions, and the solution that minimizes the transmit power at all devices is

$$(c^{(t)})_{\mathsf{opt}} = \min\left\{\frac{\sqrt{N_0}(2\zeta)^{-\frac{1}{4}}}{\sqrt{\gamma^{(t)}}}\left(1-\frac{\mu}{L}\right)^{-t/4}, \sqrt{P}\min_k \frac{h_k^{(t)}}{D_k G_k^{(t)}}\right\} \quad (45)$$

$$(\sigma_k^{(t)})_{\mathsf{opt}} = 0, \quad (46)$$

  where the value of parameter ζ can be obtained by bisection to satisfy the constraint

$$\sum_{t=1}^{T}(\sqrt{2}\gamma^{(t)})^2 \min\left\{\frac{(1-\mu/L)^{-t/2}}{\sqrt{2\zeta}\gamma^{(t)}}, \frac{P}{N_0}\min_k\left(\frac{h_k^{(t)}}{D_k G_k^{(t)}}\right)^2\right\} = \mathcal{R}_{\mathsf{dp}}(\epsilon,\delta). \quad (47)$$

  In this case, all the transmitted powers need to be scaled down in order to leverage the channel noise to ensure (ϵ,δ)-DP.

Proof: The proof follows via the same steps of Theorem 1 by replacing the local optimization problem in OMA with the optimization problem of the device with smallest value of the ratio $h_k^{(t)}/(D_k G_k^{(t)})$.

In a manner similar to OMA, Theorem 2 demonstrates that it is optimal for devices not to add further noise to the transmitted signals, i.e., to set $\mathbf{n}_k^{(t)} = 0$ in (16). Furthermore, under condition (44), privacy is attainable "for free" since the optimal solution coincides with that obtained when excluding the DP constraint (43b). As for OMA, increasing the size D_k of the data sets makes condition (44) less restrictive.

2) Online Optimization: With the offline results in Theorem 2, we are ready to describe a heuristic online approach for NOMA which follows the same logic as in Section III-C. In particular, at each iteration t, the edge server solves problem (43) over the interval $[t,\cdots,T]$ of current and future time instants by using estimated parameters $\{h_k^{(t)}, \gamma^{(t)}, G_k^{(t)}\}$, and imposing the residual DP constraint for each device. We note that the optimization problem for NOMA is solved at edge server with the known values of $\{D_k\}$.

To detail the procedure summarized in Algorithm 2, channels are predicted as in (32). Parameter $\widehat{\gamma}$ is set through the clipped per-sample gradient (35). Finally, estimates $\{\widehat{G}_k^{(t)}\}$ are obtained by using the received signal of the last iteration as described in OMA, but averaged with the number of global data set, which is given as

$$\widehat{G}_k^{(t)} = \begin{cases} \|\mathbf{y}^{(t-1)}\|/(c^{(t-1)}D_{\mathsf{tot}}), & t>1,\forall k, \\ \widehat{\gamma}, & t=1,\forall k, \end{cases} \quad (48a)$$

$$\widehat{G}_k^{(t')} = \widehat{G}_k^{(t)}, \forall\ t'>t, \forall k. \quad (48b)$$

One last issue to consider is that the optimized $c^{(t)}$ may violate the power constraint due to the use of estimated parameters. We hence modify the clipped gradient transmission as

(Clipped gradient transmission in NOMA)

$$\mathbf{x}_k^{(t)} = \min\left\{1, \frac{\sqrt{P}h_k^{(t)}}{c^{(t)}D_k\|\overline{\nabla F}_k\|}\right\}\frac{c^{(t)}}{h_k^{(t)}}D_k\overline{\nabla F}_k\left(\mathbf{w}^{(t)}\right). \quad (49)$$

As for NOMA, we make no claims of optimality, and we test the performance of the proposed online scheme via numerical results in the next section.

Algorithm 2 Online Scheme for NOMA

Input: DP level $\mathcal{R}_{\mathsf{dp}}$, channel noise $\sqrt{N_0}$, channel correlation ρ, clipping threshold $\gamma^{(t)} = \widehat{\gamma}$
Initialize: Privacy loss $\mathcal{L}_g^{(0)} = 0$.
For each iteration: $t = 1,\cdots,T$
  For edge server:
    Compute predictors $\{\widehat{h}_k^{(t')}, \widehat{G}_k^{(t')}\}$ via (32) and (48) for $t' \in [t,\cdots,T]$ with $\widehat{h}_k^{(t)} = h_k^{(t)}$
    Apply Theorem 2 over the time interval $[t,\cdots,T]$ with $\{h_k^{(t')} \leftarrow \widehat{h}_k^{(t')},\ G_k^{(t')} \leftarrow \widehat{G}_k^{(t')},\ \gamma^{(t)} \leftarrow \widehat{\gamma}\}$ and residual DP constraint $\mathcal{R}_{\mathsf{dp}} - \mathcal{L}_g^{(t-1)}$
    Broadcast optimized scaling factor $c^{(t)}$ to devices
    Update privacy loss as $\mathcal{L}_g^{(t)} = \mathcal{L}_g^{(t-1)} + \frac{(\sqrt{2}c^{(t)}\gamma^{(t)})^2}{N_0}$
  end
  For each device: $k = 1,\cdots,K$
    Update local model by (35)
    Receive $c^{(t)}$ and apply it to transmit (49).
  end
end

V. Numerical Results

In this section, we evaluate the performance of the proposed schemes in order to gain insights into the impact of the DP constraints and into the benefits of adaptive power allocation. We first consider a randomly generated synthetic dataset with $D_{\mathsf{tot}} = 10000$ pairs $(\mathbf{u}, v)$, where the covariates $\mathbf{u} \in \mathbb{R}^{10}$ are drawn i.i.d. as $\mathcal{N}(0,\mathbf{I})$ and the label v for each vector $\mathbf{u}$ is obtained as $v = u(2) + 3u(5) + 0.2z_o$, where $u(d)$ is the d-th entry in vector $\mathbf{u}$ and the observation noises $z_o \sim \mathcal{N}(0,1)$ are i.i.d. across the samples [28]. Unless stated otherwise, the training samples are evenly distributed across the $K = 10$

devices, so that the size of local data set is $D_k = 1000$ for all k. We consider ridge regression with the sample-wise loss function $f(\mathbf{w};\mathbf{u},v) = 0.5\|\mathbf{w}^\mathsf{T}\mathbf{u} - v\|^2$ and the regularization function $R(\mathbf{w}) = \|\mathbf{w}\|^2$ with $\lambda = 5\times 10^{-5}$. The PL parameter μ and smoothness parameter L are computed as the smallest and largest eigenvalues of the data Gramian matrix $\mathbf{U}^\mathsf{T}\mathbf{U}/D_{\text{tot}} + 2\lambda\mathbf{I}$, where $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_{D_{\text{tot}}}]^\mathsf{T}$ is data matrix of the data set. The initial value for $\mathbf{w}$ is set as an all-zero vector. We note that the (unique) optimal solution to the joint learning problem (3) is $\mathbf{w}^* = (\mathbf{U}^\mathsf{T}\mathbf{U} + 2D_{\text{tot}}\lambda\mathbf{I})^{-1}\mathbf{U}^\mathsf{T}\mathbf{v}$, where $\mathbf{v} = [v_1, \cdots, v_{D_{\text{tot}}}]^\mathsf{T}$ is label vector. We will also consider experiments with the MNIST data set at the end of this section.

Unless stated otherwise, the maximum SNR defined in (10) is set to $\mathsf{SNR}_{\max} = 30$ dB, and we consider the availability of 30 communication blocks. Note that this implies $T = 30/K = 3$ iterations per device for OMA and $T = 30$ iterations for NOMA. Furthermore, the default DP settings are $\epsilon = 20$ and $\delta = 0.01$.

As a benchmark, we consider a scheme that divides up the DP constraint equally across all iterations, i.e., it requires $(\sqrt{2}h_k^{(t)}\alpha_k^{(t)}\gamma^{(t)}/m_k^{(t)})^2 < \mathcal{R}_{\mathsf{dp}}/T$ for all $t = 1, \cdots, T$ in lieu of constraint (26b) and similarly for the constraint (43b). This yields

(Static PA in OMA)

$$\alpha_k^{(t)} = \min\left\{\sqrt{\frac{N_o\mathcal{R}_{\mathsf{dp}}(\epsilon,\delta)}{2T(h_k^{(t)}\gamma^{(t)})^2}}, \frac{\sqrt{P}}{D_k G_k^{(t)}}\right\}, \tag{50}$$

(Static PA in NOMA)

$$c^{(t)} = \min\left\{\sqrt{\frac{N_o\mathcal{R}_{\mathsf{dp}}(\epsilon,\delta)}{2T(\gamma^{(t)})^2}}, \sqrt{P}\min_k \frac{h_k^{(t)}}{D_k G_k^{(t)}}\right\}. \tag{51}$$

Another benchmark is set by the scheme that doe not impose the DP constraint (26b) and (43b). We adopt the normalized optimality gap $[F(\mathbf{w}^{T+1}) - F(\mathbf{w}^*)]/F(\mathbf{w}^*)$ as performance metric, and the offline results are averaged over 1000 channel realizations while online results are averaged over 100 channel realizations.

A. Offline Optimization

We now focus on offline optimization by applying the optimal adaptive PA strategies in Theorems 1 and 2. For the channel model in (31), we set $\kappa = 10$, and the channel correlation parameter is set as $\rho = 1$, since this parameter has no discernible effect on the performance of offline strategies. We use the simple upper bounds $\gamma^{(t)} = 2W\max_{(\mathbf{u},v)\in\mathcal{D}} L(\mathbf{u},v)$ and $G_k^{(t)} = 2WL_k$, where $W \geq \|\mathbf{w}\|$ is a bound on the norm $\|\mathbf{w}\|$ (which can be in practice ensured via convex projection and is set to $W = 3.2$ in our results); and $L(\mathbf{u},v)$ and L_k are the Lipschitz smoothness constants of functions $f(\mathbf{w};\mathbf{u}_n,v_n)$ and $F_k(\mathbf{w})$, respectively.

In Fig. 3, we plot the normalized optimality gap as a function of the privacy level ϵ. In the considered range of ϵ, NOMA with either adaptive or static power allocation (PA) is seen to achieve better performance than OMA. Furthermore, adaptive PA achieves a significant performance gain over static PA under stringent DP constraints, while the performance

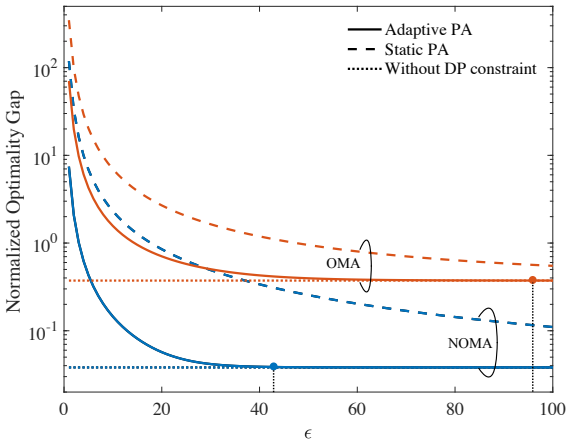


Figure 3. Optimality gap versus DP privacy level ϵ (for $\delta = 0.01$) for different power allocation (PA) schemes and for the scheme without DP constraint ($\mathsf{SNR}_{\max} = 30$ dB).

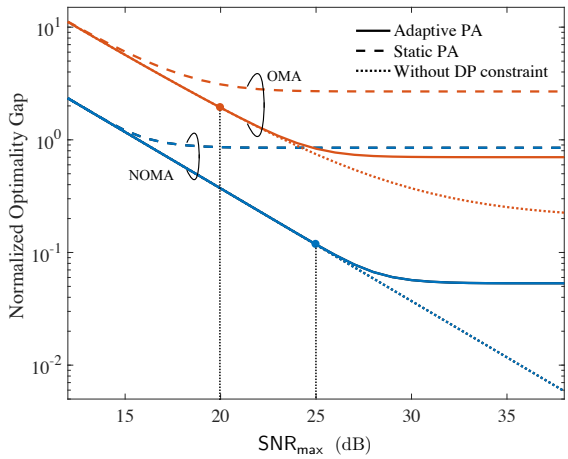


Figure 4. Optimality gap versus $\mathsf{SNR}_{\max}$ for different power allocation (PA) schemes and for the scheme without DP constraint ($\epsilon = 20$, $\delta = 0.01$).

advantage of adaptive PA decreases as the DP constraint is relaxed, i.e., for larger values of ϵ. The figure also shows the threshold values of ϵ beyond which the privacy "for free" conditions (27) and (44) are satisfied.

We now study the impact of the $\mathsf{SNR}_{\max}$ (10) in Fig. 4. The normalized optimality gap of all schemes is seen to decrease with the SNR until the DP requirement becomes the performance bottleneck. While NOMA is confirmed to be generally advantageous over OMA, OMA with optimal PA can perform better than NOMA with static PA, which emphasizes the importance of PA optimization, particularly in the high-SNR regime. In a manner analogous to Fig. 3, the plot also marks the maximum SNR levels for which the privacy "for free" condition (27) and (44) are satisfied.

Fig. 5 plots the normalized optimality gap versus a measure of the heterogeneity of the data sets. To this end, one of the devices is allocated a lager value D_k, while the remaining data points are evenly distributed to the other devices. The ratio $\max_k D_k/D_{\text{tot}}$ varies from 0.1 (uniformly distributed) to 0.95 (highly skewed). Increasing data set heterogeneity generally affects negatively all schemes, even in the absence of

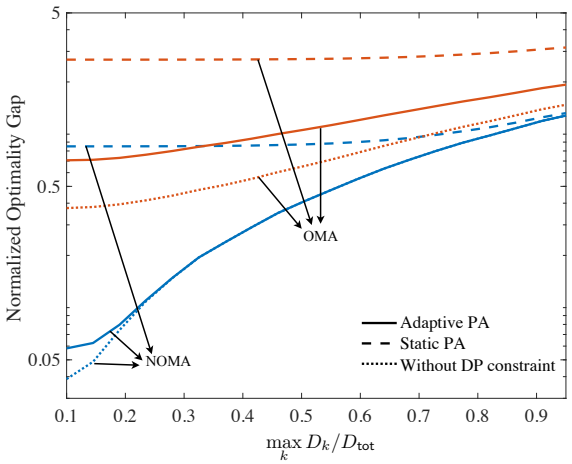


Figure 5. Optimality gap versus data set heterogeneity parameter $\max_k D_k/D_{\text{tot}}$ for different power allocation (PA) schemes and the scheme without DP constraint ($\epsilon = 20$, $\delta = 0.01$, $\mathsf{SNR}_{\max} = 30$ dB).

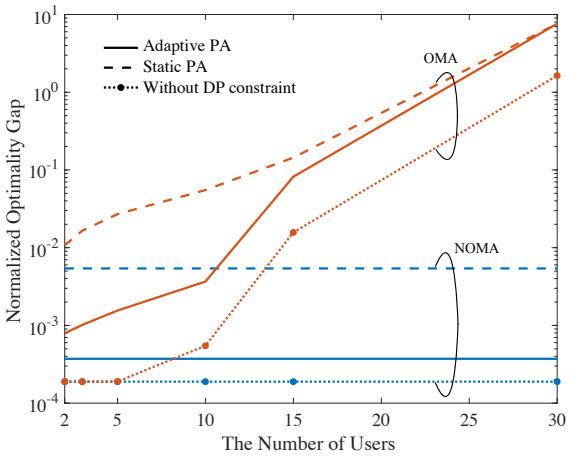


Figure 7. Optimality gap versus the number of user K for different power allocation (PA) schemes and for the scheme without DP constraint ($\epsilon = 20$, $\delta = 0.01$, $\mathsf{SNR}_{\max} = 30$ dB, $\kappa = 5$, $\rho = 0$).

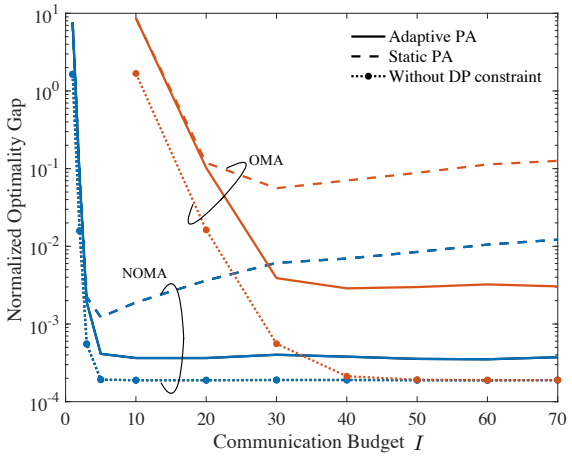


Figure 6. Optimality gap versus communication budget I for different power allocation (PA) schemes and the scheme without DP constraint ($\epsilon = 20$, $\delta = 0.01$, $\mathsf{SNR}_{\max} = 30$ dB, $\kappa = 5$, $\rho = 0$).

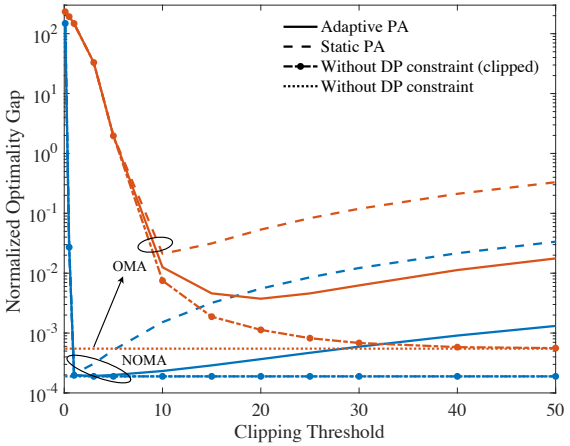


Figure 8. Optimality gap versus the clipping threshold $\widehat{\gamma}$ for different power allocation (PA) schemes and for the schemes without DP constraint ($\epsilon = 20$, $\delta = 0.01$, $\mathsf{SNR}_{\max} = 30$ dB, $T = 30$).

privacy constraints. Nevertheless, the heterogeneity of D_k has a stronger impact for NOMA than for OMA due to gradient alignment condition (36). In particular, for NOMA, the power constraint becomes the performance bottleneck as the ratio $\max D_k/D_{\text{tot}}$ increases, and the performance of adaptive PA converges first to that without the DP constraint and then to that of static PA.

B. Online Optimization

We now turn to the heuristic online optimization methods proposed in Algorithms 1 and 2. For the channel model in (31), we set $\kappa = 5$ and $\rho = 0$. Note that channel prediction is possible due to the non-zero Rician factor. The maximum value of $\|\mathbf{w}\|$ is set as $W = 10$, which is ensured by convex projection. Unless stated otherwise, we set the clipping threshold as $\widehat{\gamma} = 20$.

In Fig. 6, we study the impact of the communication budget in terms of number of communication blocks I. With conventional static PA, there exists an optimal communication budget under privacy constraints. This is because more communication blocks may cause an increase in privacy loss (see also [25]). In contrast, increasing the communication budget always benefits adaptive PA, which is able to properly allocate power across the communication blocks. Furthermore, without DP constraint, the performance of OMA converges to that of NOMA when the communication budget I is large; while, under privacy constraints, NOMA retains performance advantages even with a large I.

Fig. 7 plots the normalized optimality gap versus the number of users. It shows that increasing the number of users has a negative effect on OMA, but it causes no harm to NOMA. This emphasizes the spectral efficiency of NOMA in wireless edge learning. Furthermore, under OMA, a larger number of users implies fewer iterations, and thus less information leakage of each user, decreasing the performances gain of adaptive PA. Specifically, when $K = 30$, a simple iteration $T = 1$ is carried out by OMA, and adaptive PA is equivalent to static PA.

We now study the impact of the clipping threshold $\widehat{\gamma}$ for the gradient in Fig. 8. To show the impact of clipping, we also plot the performance with clipped local updates without

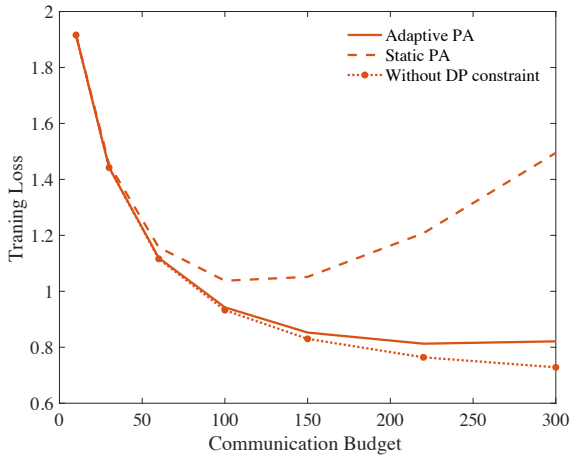
the DP constraint for both OMA and NOMA. Without DP constraint, the larger clipping threshold incurs a smaller distortion of the gradients, which benefits the learning performance. However, under the DP constraint, increasing the clipping threshold beyond a given value degrades the performance, since ensuring privacy requires a more pronounced scaling down of the transmitted signals. This indicates the importance of selecting a threshold $\widehat{\gamma}$ that strikes a balance between learning performance and privacy.

C. MNIST Data Set

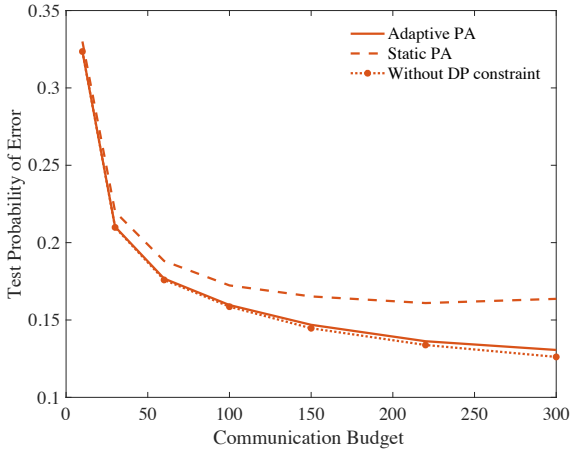
We now consider the problem of classification on the MNIST data set via multinomial logistic regression with quadratic regularization. Accordingly, the global loss function is given as the regularized cross-entropy loss

$$F(\mathbf{w}) = \frac{1}{D_{\text{tot}}} \sum_{(\mathbf{u},v)\in\mathcal{D}} \sum_{c=1}^{C} \mathbf{1}\{v=c\} \log \frac{\exp(\mathbf{w}_c^{\mathsf{T}}\mathbf{u})}{\sum_{j=1}^{C}\exp(\mathbf{w}_j^{\mathsf{T}}\mathbf{u})} + \lambda \sum_{c=1}^{C} \|\mathbf{w}_c\|^2,$$

where $C = 10$ represents the total number of classes of handwritten digits; $\mathbf{u}$ is data image extended to include a bias term; and the model parameter $\mathbf{w}$, with dimension 7650, is comprised of the per-class vectors $\{\mathbf{w}_c\}_{c=1}^{C}$. We set $\lambda = 0.01$, the maximum value of $\|\mathbf{w}\|$ to $W = 10$, the clipping threshold as $\widehat{\gamma} = 40$, and $\mathsf{SNR}_{\max} = 13$ dB. The smoothness parameter L and strongly convex parameter μ are treated as hyper-parameter and selected via validation as $\mu = 0.3$ and $L = 2.5$. For $\epsilon = 5$ and $\delta = 0.01$, Fig. 9 plots the training cross-entropy loss and the probability of error on the test set versus the value of communication budget I for OMA. Adaptive PA is seen to significantly outperform static PA both in terms of training loss and test error. Similar results can be obtained for NOMA.



(a) Training cross-entropy loss versus communication budget I.



(b) Test probability of error versus communication budget I.

Figure 9. Training loss and test error for different power allocation (PA) schemes on the MNIST dat set for OMA ($\epsilon = 5$, $\delta = 0.01$, $\mathsf{SNR}_{\max} = 13$ dB, $\kappa = 5$, $\rho = 0$).

VI. Conclusions

In this paper, we have considered differentially private wireless federated learning via the direct, uncoded, transmission of gradient from devices to edge server. The proposed approach is based on adaptive PA schemes that are optimized to minimize the learning optimality gap under privacy and power constraints. First, offline optimization problems are separately formulated for OMA and NOMA, which are converted to convex programs. The optimal PA, obtained in closed form, adapts the power along the iterations, outperforming static PA assumed in prior works. Furthermore, a heuristic online approach is proposed that leverages iterative one-step-ahead optimization based on the offline result and predicted CSI.

The analysis in this paper proved that privacy can be obtained "for free", that is without affecting the learning performance, as long as the privacy constraint level is below a threshold that decreases with the SNR. Our analytical results also demonstrate that it is generally suboptimal to devote part of the transmitted power to actively add noise to the local updates. This is unlike the standard scenario with ideal communication, in which adding noise is essential to ensure DP constraints. Via numerical results, we have finally shown that techniques that leverage over-the-air computing provide significant benefits over conventional OMA protocols under DP constraints. This is not a prior evident, since, with NOMA, devices transmit more frequently, and hence may leak more information.

We note that the power control policy based on channel inversion for all the devices was proven to be suboptimal in the scenario of over-the-air computing without DP constraint. In fact, channel inversion can incur noise amplification by adapting the power to the device with worst channel condition [11], [38]. However, this may not be the case under DP constraints since noise amplification benefits privacy. As a possible extension of the current work, it would be interesting to study the optimization of the threshold for channel inversion so as to maximize the learning performance under privacy and power constraints. As directions for future work, the threat model could also include "honest-but-curious" edge devices, which would generally incur larger DP loss. The study could be further generalized to other network topologies including multi-hop device-to-device (D2D) networks. Another interesting direction is to consider the implementation of digital transmission where quantization introduces additional privacy

preserving mechanism on top of the channel noise. It would also be interesting to investigate the effect of clipping in terms of convergence as in [44], [45], and to address the convergence properties of the proposed online scheme.

APPENDIX

A. Proof of Lemma 1

To start, we denote as $\mathbf{y}_k = [\mathbf{y}_k^{(1)}, \cdots, \mathbf{y}_k^{(T)}]$ the T successive received signals from device k, and $m_k^{(t)} = \sqrt{(h_k^{(t)}\alpha_k^{(t)}\sigma_k^{(t)})^2 + N_0}$ is the standard deviation of the effective noise in $\mathbf{y}_k^{(t)}$. According to the definition of DP loss given in (12), for the k-th device, the privacy loss after T iterations can be represented as

$$\begin{aligned}
\mathcal{L}_{\mathcal{D},\mathcal{D}'}(\mathbf{y}_k) &= \ln\left(\prod_{t=1}^{T}\frac{P\left[\mathbf{y}_k^{(t)}|\mathbf{y}_k^{(t-1)},\cdots,\mathbf{y}_k^{(1)},\mathcal{D}_k\right]}{P\left[\mathbf{y}_k^{(t)}|\mathbf{y}_k^{(t-1)},\cdots,\mathbf{y}_k^{(1)},\mathcal{D}'_k\right]}\right)\\
&= \sum_{t=1}^{T}\ln\left(\frac{P\left[\mathbf{y}_k^{(t)}|\mathbf{y}_k^{(t-1)},\cdots,\mathbf{y}_k^{(1)},\mathcal{D}_k\right]}{P\left[\mathbf{y}_k^{(t)}|\mathbf{y}_k^{(t-1)},\cdots,\mathbf{y}_k^{(1)},\mathcal{D}'_k\right]}\right)\\
&= \sum_{t=1}^{T}\ln\left(\frac{\exp\left(-\frac{\|\mathbf{y}_k^{(t)}-h_k^{(t)}\alpha_k^{(t)}D_k\nabla F_k(\mathbf{w}^{(t)};\mathcal{D}_k)\|^2}{2(m_k^{(t)})^2}\right)}{\exp\left(-\frac{\|\mathbf{y}_k^{(t)}-h_k^{(t)}\alpha_k^{(t)}D_k\nabla F_k(\mathbf{w}^{(t)};\mathcal{D}'_k)\|^2}{2(m_k^{(t)})^2}\right)}\right)\\
&= \sum_{t=1}^{T}\ln\left(\frac{\exp\left(-\frac{\|\mathbf{r}_k^{(t)}\|^2}{2(m_k^{(t)})^2}\right)}{\exp\left(-\frac{\|\mathbf{r}_k^{(t)}+\mathbf{v}_k^{(t)}\|^2}{2(m_k^{(t)})^2}\right)}\right),
\end{aligned}$$

where $\mathbf{r}_k^{(t)} \sim \mathcal{N}(0,(m_k^{(t)})^2\mathbf{I})$ represents the effective noise, and we set

$$\mathbf{v}_k^{(t)} = h_k^{(t)}\alpha_k^{(t)}\Big[\sum_{(\mathbf{u},v)\in\mathcal{D}_k}\nabla f\left(\mathbf{w}^{(t)};\mathbf{u},v\right) - \sum_{(\mathbf{u},v)\in\mathcal{D}'_k}\nabla f\left(\mathbf{w}^{(t)};\mathbf{u},v\right)\Big]$$

with $\|\mathbf{v}_k^{(t)}\| = \Delta_k^{(t)}$. Following [9, Appendix A], we can then bound privacy violation probability

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T}\frac{2(\mathbf{r}_k^{(t)})^{\mathsf{T}}\mathbf{v}_k^{(t)}+\|\mathbf{v}_k^{(t)}\|^2}{2(m_k^{(t)})^2}\right|>\epsilon\right)\\
&\overset{(a)}{\leq}\Pr\left(\left|\sum_{t=1}^{T}\frac{(\mathbf{r}_k^{(t)})^{\mathsf{T}}\mathbf{v}_k^{(t)}}{(m_k^{(t)})^2}\right|>\epsilon-\sum_{t=1}^{T}\frac{\|\mathbf{v}_k^{(t)}\|^2}{2(m_k^{(t)})^2}\right)\\
&=2\Pr\left(\sum_{t=1}^{T}\frac{(\mathbf{r}_k^{(t)})^{\mathsf{T}}\mathbf{v}_k^{(t)}}{(m_k^{(t)})^2}>\epsilon-\sum_{t=1}^{T}\frac{\|\mathbf{v}_k^{(t)}\|^2}{2(m_k^{(t)})^2}\right)\\
&\overset{(b)}{\leq}2\frac{\sqrt{\sum_{t=1}^{T}\left(\frac{\Delta_k^{(t)}}{m_k^{(t)}}\right)^2}}{\sqrt{2\pi}\left[\epsilon-\sum_{t=1}^{T}\frac{1}{2}\left(\frac{\Delta_k^{(t)}}{m_k^{(t)}}\right)^2\right]}\\
&\qquad\times\exp\left(-\frac{\left[\epsilon-\sum_{t=1}^{T}\frac{1}{2}\left(\frac{\Delta_k^{(t)}}{m_k^{(t)}}\right)^2\right]^2}{2\sum_{t=1}^{T}\left(\frac{\Delta_k^{(t)}}{m_k^{(t)}}\right)^2}\right),
\end{aligned}\tag{52}$$

where (a) is obtained by using $\Pr(X<-\epsilon-b)\leq\Pr(X<-\epsilon+b)$ for an arbitrary $b\geq 0$, and (b) comes from the following bound on the tail probability of Gaussian distribution $X\sim\mathcal{N}(0,\sigma^2)$: $\Pr(X>s)=\frac{1}{\sigma\sqrt{2\pi}}\int_s^\infty\exp(-\frac{x^2}{2\sigma^2})dx\leq\frac{1}{\sigma\sqrt{2\pi}}\int_s^\infty\frac{x}{s}\exp(-\frac{x^2}{2\sigma^2})dx=\frac{\sigma}{s\sqrt{2\pi}}\exp\left(-\frac{s^2}{2\sigma^2}\right)$.

Letting $q=\frac{\epsilon-\sum_{t=1}^{T}\frac{1}{2}(\Delta_k^{(t)}/m_k^{(t)})^2}{\sqrt{2\sum_{t=1}^{T}(\Delta_k^{(t)}/m_k^{(t)})^2}}$ and using (52), the DP condition is implied by the inequality

$$\Pr(|\mathcal{L}_{\mathcal{D},\mathcal{D}'}(\mathbf{y}_k)|>\epsilon)\leq\frac{1}{q\sqrt{\pi}}e^{-q^2}<\delta.\tag{53}$$

Finally, defining the function $\mathcal{C}(x)=\sqrt{\pi}xe^{x^2}$ and utilizing its monotonicity yields the desired result.

B. Proof of Lemma 2

Under Assumption 1, we have the following equality

$$\begin{aligned}
F\big(\mathbf{w}^{(t)}\big)&\leq F\big(\mathbf{w}^{(t-1)}\big)+\Big[\nabla F\big(\mathbf{w}^{(t-1)}\big)\Big]^{\mathsf{T}}\Big[\mathbf{w}^{(t)}-\mathbf{w}^{(t-1)}\Big]\\
&\quad+\frac{L}{2}\Big\|\mathbf{w}^{(t)}-\mathbf{w}^{(t-1)}\Big\|^2\\
&=F\big(\mathbf{w}^{(t-1)}\big)-\eta\Big[\nabla F\big(\mathbf{w}^{(t-1)}\big)\Big]^{\mathsf{T}}\times\Big[\nabla F\big(\mathbf{w}^{(t-1)}\big)\\
&\quad+\frac{1}{D_{\mathsf{tot}}}\sum_{k=1}^{K}\Big[\mathbf{n}_k^{(t-1)}+\Big(h_k^{(t-1)}\alpha_k^{(t-1)}\Big)^{-1}\mathbf{z}_k^{(t-1)}\Big]\Big]\\
&\quad+\frac{L\eta^2}{2}\Big\|\nabla F\big(\mathbf{w}^{(t-1)}\big)+\frac{1}{D_{\mathsf{tot}}}\sum_{k=1}^{K}\Big[\mathbf{n}_k^{(t-1)}+\\
&\qquad\Big(h_k^{(t-1)}\alpha_k^{(t-1)}\Big)^{-1}\mathbf{z}_k^{(t-1)}\Big]\Big\|^2.
\end{aligned}$$

By taking the expectation over the additive noise on both sides of the above inequality, we obtain

$$\begin{aligned}
\mathsf{E}\Big[F\big(\mathbf{w}^{(t)}\big)\Big]&\leq F\big(\mathbf{w}^{(t-1)}\big)-\eta\Big[\Big(1-\frac{L\eta}{2}\Big)\Big\|\nabla F\big(\mathbf{w}^{(t-1)}\big)\Big\|^2\Big]\\
&\quad+\frac{L\eta^2 d}{2D_{\mathsf{tot}}^2}\sum_{k=1}^{K}(\sigma_k^{(t-1)})^2+\left(\frac{\sqrt{N_0}}{h_k^{(t-1)}\alpha_k^{(t-1)}}\right)^2\\
&=F\big(\mathbf{w}^{(t-1)}\big)-\frac{1}{2L}\Big\|\nabla F\big(\mathbf{w}^{(t-1)}\big)\Big\|^2\\
&\quad+\frac{d}{2LD_{\mathsf{tot}}^2}\sum_{k=1}^{K}\left(\frac{m_k^{(t-1)}}{h_k^{(t-1)}\alpha_k^{(t-1)}}\right)^2,
\end{aligned}$$

where the equality follows by Lemma 1 and by setting $\eta=1/L$.

Subtracting the optimal value F^* at both sides yields

$$\begin{aligned}
&\mathsf{E}\Big[F\big(\mathbf{w}^{(t)}\big)\Big]-F^*\\
&\leq F\big(\mathbf{w}^{(t-1)}\big)-F^*-\frac{1}{2L}\Big\|\nabla F\big(\mathbf{w}^{(t-1)}\big)\Big\|^2\\
&\quad+\frac{d}{2LD_{\mathsf{tot}}^2}\sum_{k=1}^{K}\left(\frac{m_k^{(t-1)}}{h_k^{(t-1)}\alpha_k^{(t-1)}}\right)^2\\
&\leq\Big(1-\frac{\mu}{L}\Big)\Big(F\big(\mathbf{w}^{(t-1)}\big)-F^*\Big)+\frac{d}{2LD_{\mathsf{tot}}^2}\\
&\quad\times\sum_{k=1}^{K}\left(\frac{m_k^{(t-1)}}{h_k^{(t-1)}\alpha_k^{(t-1)}}\right)^2,
\end{aligned}\tag{54}$$

where the last step follows from Assumption 2. Then, the desired result yields by applying above inequality repeatedly through T iterations and taking expectation over all the additive noises.

C. Proof of Theorem 1

We start by making the change of variables

$$a_k^{(t)} = (\sigma_k^{(t)})^2 + N_0/(h_k^{(t)}\alpha_k^{(t)})^2, \quad b_k^{(t)} = (\alpha_k^{(t)})^{-2}, \quad (55)$$

so that the original variables can be written as

$$(\sigma_k^{(t)})^2 = a_k^{(t)} - (\sqrt{N_0}/h_k^{(t)})^2 b_k^{(t)} \geq 0,\ (\alpha_k^{(t)})^2 = 1/b_k^{(t)} > 0. \quad (56)$$

By including the constraints (56), we now obtain the equivalent local problem (**OMA Local Opt. 2**)

$$\begin{aligned}
\min_{\{a_k^{(t)}, b_k^{(t)}\}} \quad & \sum_{t=1}^{T} \left(1-\frac{\mu}{L}\right)^{-t} a_k^{(t)} \\
\text{s.t.} \quad & \sum_{t=1}^{T} (\sqrt{2}\gamma^{(t)})^2/a_k^{(t)} \leq \mathcal{R}_{\mathsf{dp}}, \\
& (D_k G_k^{(t)})^2 + d a_k^{(t)} - \left[d(\sqrt{N_0}/h_k^{(t)})^2 + P\right] b_k^{(t)} \leq 0, \quad \forall t, \\
& a_k^{(t)} - (\sqrt{N_0}/h_k^{(t)})^2 b_k^{(t)} \geq 0, \quad \forall t, \\
& a_k^{(t)} \geq 0, \quad b_k^{(t)} \geq 0, \qquad \forall t,
\end{aligned}$$

which is a convex optimization problem. To solve it, the partial Lagrange function is defined as

$$\begin{aligned}
\mathcal{L} = & \sum_{t=1}^{T} \left(1-\frac{\mu}{L}\right)^{-t} a_k^{(t)} + \zeta \left(\sum_{t=1}^{T} \frac{(\sqrt{2}\gamma^{(t)})^2}{a_k^{(t)}} - \mathcal{R}_{\mathsf{dp}} \right) \\
& + \sum_{t=1}^{T} \xi^{(t)} \left(\left(\sqrt{N_0}/h_k^{(t)}\right)^2 b_k^{(t)} - a_k^{(t)} \right) \\
& + \sum_{t=1}^{T} \beta^{(t)} \Big((D_k G_k^{(t)})^2 + d a_k^{(t)} \\
& \qquad - \left[d\big(\sqrt{N_0}/h_k^{(t)}\big)^2 + P\right] b_k^{(t)} \Big), \quad (57)
\end{aligned}$$

where $\zeta \geq 0$, $\beta^{(t)} \geq 0$, and $\xi^{(t)} \geq 0$ are the Lagrange multipliers associated respectively with the DP constraint, transmit power constraints and non-negative parameter constraints. Then applying the KKT conditions leads to the following necessary and sufficient conditions

$$\begin{aligned}
& \frac{\partial \mathcal{L}}{\partial (a_k^{(t)})_{\mathsf{opt}}} = \left(1-\frac{\mu}{L}\right)^{-t} - \zeta_{\mathsf{opt}} (\sqrt{2}\gamma^{(t)})^2 \left((a_k^{(t)})_{\mathsf{opt}}\right)^{-2} \\
& \qquad + (\beta^{(t)})_{\mathsf{opt}} d - (\xi^{(t)})_{\mathsf{opt}} = 0, \quad (58a) \\
& \frac{\partial \mathcal{L}}{\partial (b_k^{(t)})_{\mathsf{opt}}} = -(\beta^{(t)})_{\mathsf{opt}} \left[d(\sqrt{N_0}/h_k^{(t)})^2 + P\right] \\
& \qquad + (\xi^{(t)})_{\mathsf{opt}} (\sqrt{N_0}/h_k^{(t)})^2 = 0, \quad (58b) \\
& \zeta_{\mathsf{opt}} \left(\sum_{t=1}^{T} \frac{(\sqrt{2}\gamma^{(t)})^2}{(a_k^{(t)})_{\mathsf{opt}}} - \mathcal{R}_{\mathsf{dp}} \right) = 0, \quad (58c) \\
& (\beta^{(t)})_{\mathsf{opt}} \Big((D_k G_k^{(t)})^2 + d (a_k^{(t)})_{\mathsf{opt}} - \left[d\big(\sqrt{N_0}/h_k^{(t)}\big)^2 + P\right] \\
& \qquad \times (b_k^{(t)})_{\mathsf{opt}} \Big) = 0, \quad (58d) \\
& (\xi^{(t)})_{\mathsf{opt}} \left[\left(\sqrt{N_0}/h_k^{(t)}\right)^2 (b_k^{(t)})_{\mathsf{opt}} - (a_k^{(t)})_{\mathsf{opt}} \right] = 0, \quad (58e) \\
& \sum_{t=1}^{T} \frac{(\sqrt{2}\gamma^{(t)})^2}{(a_k^{(t)})_{\mathsf{opt}}} - \mathcal{R}_{\mathsf{dp}} \leq 0, \quad (58f) \\
& (D_k G_k^{(t)})^2 + d (a_k^{(t)})_{\mathsf{opt}} - \left[d\big(\sqrt{N_0}/h_k^{(t)}\big)^2 + P\right] (b_k^{(t)})_{\mathsf{opt}} \\
& \qquad \leq 0, \quad (58g) \\
& \left(\sqrt{N_0}/h_k^{(t)}\right)^2 (b_k^{(t)})_{\mathsf{opt}} - (a_k^{(t)})_{\mathsf{opt}} \leq 0. \quad (58h)
\end{aligned}$$

According to (58b), we have the following equality for the optimal solutions $(\beta^{(t)})_{\mathsf{opt}}$ and $(\xi^{(t)})_{\mathsf{opt}}$

$$(\xi^{(t)})_{\mathsf{opt}} = \frac{d(\sqrt{N_0}/h_k^{(t)})^2 + P}{(\sqrt{N_0}/h_k^{(t)})^2} (\beta^{(t)})_{\mathsf{opt}}. \quad (59)$$

Plugging the above result into (58a) and (58e), respectively, we obtain

$$\begin{aligned}
& \left(1-\frac{\mu}{L}\right)^{-t} - \zeta_{\mathsf{opt}} (\sqrt{2}\gamma^{(t)})^2 \left((a_k^{(t)})_{\mathsf{opt}}\right)^{-2} \\
& \qquad - (\beta^{(t)})_{\mathsf{opt}} \frac{P(h_k^{(t)})^2}{(\sqrt{N_0})^2} = 0 \quad (60) \\
& (\beta^{(t)})_{\mathsf{opt}} \frac{d(\sqrt{N_0}/h_k^{(t)})^2 + P}{(\sqrt{N_0}/h_k^{(t)})^2} \Big(\left(\sqrt{N_0}/h_k^{(t)}\right)^2 (b_k^{(t)})_{\mathsf{opt}} \\
& \qquad - (a_k^{(t)})_{\mathsf{opt}} \Big) = 0. \quad (61)
\end{aligned}$$

Combining (61) and (58d), we get the following equation

$$(\beta^{(t)})_{\mathsf{opt}} \left((D_k G_k^{(t)})^2 - P(h_k^{(t)}/\sqrt{N_0})^2 (a_k^{(t)})_{\mathsf{opt}} \right) = 0. \quad (62)$$

Constraints (58g) and (58h) define the minimum and maximum values of $(b_k^{(t)})_{\mathsf{opt}}$ in terms of $(a_k^{(t)})_{\mathsf{opt}}$. Accordingly, the minimum value of $(b_k^{(t)})_{\mathsf{opt}}$ should be no larger than that of the maximum value, which yields the following lower bound on $(a_k^{(t)})_{\mathsf{opt}}$:

$$(a_k^{(t)})_{\mathsf{opt}} \geq (D_k G_k^{(t)})^2 (\sqrt{N_0}/h_k^{(t)})^2 / P. \quad (63)$$

In this case, the power is fully utilized for transmitting the local gradient

Furthermore, from (62), we have the equality $(\beta^{(t)})_{\mathsf{opt}} = 0$ if $(a_k^{(t)})_{\mathsf{opt}} > (D_k G_k^{(t)})^2(\sqrt{N_0}/h_k^{(t)})^2/P$, thereby the other solution of $(a_k^{(t)})_{\mathsf{opt}}$ is obtained by solving (60) as

$$(a_k^{(t)})_{\mathsf{opt}} = \sqrt{2\zeta_{\mathsf{opt}}}(1-\mu/L)^{t/2}\gamma^{(t)}. \tag{64}$$

Combing (63) and (64), the solution of $(a_k^{(t)})_{\mathsf{opt}}$ is

$$(a_k^{(t)})_{\mathsf{opt}} = \max\Big\{\sqrt{2\zeta_{\mathsf{opt}}}(1-\mu/L)^{t/2}\gamma^{(t)}, (D_k G_k^{(t)})^2(\sqrt{N_0}/h_k^{(t)})^2/P\Big\}, \tag{65}$$

and the value of ζ_{opt} can be obtained by bisection search to satisfy the equality of (58f). Specifically, we have $\zeta_{\mathsf{opt}} = 0$ if $\sum_{t=1}^{T}(\sqrt{2}\gamma^{(t)}h_k^{(t)})^2 P/(\sqrt{N_0}D_k G_k^{(t)})^2 < \mathcal{R}_{\mathsf{dp}}$. With the value of $(a_k^{(t)})_{\mathsf{opt}}$, the solution of $(b_k^{(t)})_{\mathsf{opt}}$ can be obtained by using (58g) and (58h), which are satisfied by arbitrary value within the range

$$\frac{(D_k G_k^{(t)})^2 + d(a_k^{(t)})_{\mathsf{opt}}}{d(\sqrt{N_0}/h_k^{(t)})^2 + P} \le (b_k^{(t)})_{\mathsf{opt}} \le \frac{(h_k^{(t)})^2(a_k^{(t)})_{\mathsf{opt}}}{N_0}. \tag{66}$$

Then, the desired result in the theorem is obtained by reverting to the original variables using (56). Specifically, the optimal solution to minimize the transmit power is attained by the maximum value of $(b_k^{(t)})_{\mathsf{opt}}$ in (66).

REFERENCES

[1] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[3] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[5] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *[Online]. Available: https://arxiv.org/pdf/1912.04977.pdf*, 2019.

[6] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *IEEE Symp. Secur. and Privacy (SP)*, pp. 691–706, IEEE, 2019.

[7] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, (Denver, USA), Oct. 2015.

[8] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," *[Online]. Available: https://arxiv.org/pdf/1812.00984.pdf*, 2018.

[9] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

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

[11] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, 2019.

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

[13] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[14] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *[Online]. Available: https://arxiv.org/pdf/2003.02089.pdf*, 2020.

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

[16] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Adv. Neural Info. Proc. Syst. (NIPS)*, (Long Beach, USA), Dec. 2017.

[17] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. Intl. Conf. Mach. Learning (ICML)*, (Stockholmsmässan, Stockholm Sweden), July 2018.

[18] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 2128–2142, 2020.

[19] W.-T. Chang and R. Tandon, "Communication efficient federated learning over multiple access channels," *[Online]. Available: https://arxiv.org/pdf/2001.08737.pdf*, 2020.

[20] G. Zhu, Y. Du, D. Gunduz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *[Online]. Available: https://arxiv.org/pdf/2001.05713.pdf*, 2020.

[21] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symp. Secur. and Privacy (SP)*, pp. 3–18, IEEE, 2017.

[22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, (Vienna, Austria), Oct. 2016.

[23] N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar, "The value of collaboration in convex machine learning with differential privacy," *[Online]. Available: https://arxiv.org/pdf/1906.09679.pdf*, 2019.

[24] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," in *Adv. Neural Info. Proc. Syst. (NIPS)*, (Montreal, Canada), Dec. 2018.

[25] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, 2020.

[26] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Adv. Neural Info. Proc. Syst. (NIPS)*, (Montreal, Canada), Dec. 2018.

[27] V. Gandikota, R. K. Maity, and A. Mazumdar, "vqSGD: Vector quantized stochastic gradient descent," *[Online]. Available: https://arxiv.org/pdf/1911.07971.pdf*, 2019.

[28] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," *[Online]. Available: https://arxiv.org/pdf/2002.05151.pdf*, 2020.

[29] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," *[Online]. Available: https://arxiv.org/pdf/2004.06337.pdf*, 2020.

[30] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," *[Online]. Available: https://arxiv.org/pdf/2005.07776.pdf*, 2020.

[31] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *[Online]. Available: https://arxiv.org/pdf/1806.00582.pdf*, 2018.

[32] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, 2019.

[33] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Intl. Conf. Artif. Intell. Stat. (AISTATS)*, (Palermo, Italy), pp. 2021–2031, June 2020.

[34] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *[Online]. Available: https://arxiv.org/pdf/1812.06127.pdf*, 2018.

[35] S. Ravi, "Efficient on-device models using neural projections," in *Proc. Intl. Conf. Mach. Learning (ICML)*, (Long Beach, USA), pp. 5370–5379, June 2019.

[36] W. Debaenst, A. Feys, I. Cuiñas, M. García Sánchez, and J. Verhaevert, "RMS delay spread vs. coherence bandwidth from 5G indoor radio channel measurements at 3.5 GHz band," *Sensors*, vol. 20, no. 3, p. 750, 2020.

[37] S. Wang, K. Guan, D. He, G. Li, X. Lin, B. Ai, and Z. Zhong, "Doppler shift and coherence time of 5G vehicular channels at 3.5 GHz," in *Proc. IEEE Intl. Symp. on Antennas and Propagation & USNC-URSI National Radio Science Meeting*, (Boston, USA), pp. 2005–2006, IEEE, July 2018.

[38] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. Wireless Commun.*, 2020.

[39] A. Mahmood, M. I. Ashraf, M. Gidlund, J. Torsner, and J. Sachs, "Time synchronization in 5G wireless edge: Requirements and solutions for critical-mtc," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 45–51, 2019.

[40] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition," in *Joint European Conf. on Mach. Learn. Knowl. Discovery in Databases (ECML KDD)*, (Riva del Garda, Italy), Sep. 2016.

[41] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[42] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[43] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *Proc. Intl. Conf. Mach. Learning (ICML)*, (Lille, France), July 2015.

[44] X. Chen, Z. S. Wu, and M. Hong, "Understanding gradient clipping in private sgd: A geometric perspective," *[Online]. Available: https://arxiv.org/pdf/2006.15429.pdf*, 2020.

[45] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *Proc. Intl. Conf. Learning Representations (ICLR)*, (New Orleans, USA), May 2019.