# Private Over-the-Air Federated Learning at Band-Limited Edge

Youming Tao, Shuzhen Chen, Congwei Zhang, Di Wang, Dongxiao Yu, *Senior Member, IEEE*, Xiuzhen Cheng, *Fellow, IEEE*, Falko Dressler *Fellow, IEEE*

**Abstract**—We investigate over-the-air federated learning (OTA-FL) that exploits over-the-air computing (AirComp) to integrate communication and computation seamlessly for FL. Privacy presents a serious obstacle for OTA-FL, as it can be compromised by maliciously manipulating channel state information (CSI). Moreover, the limited band at edge hinders OTA-FL from training large-scale models. It remains open how to enable a multitude of devices with constrained resources and sensitive data to collaboratively train a global model at band-limited edge. To tackle this, we design a novel algorithm PROBE building upon a lightweight over-the-air gradients aggregation rule PB-O-GAR. Specifically, PB-O-GAR combines a random sparsification-like dimension reduction with Gaussian perturbation to provide rigorous privacy and band-adapted communication. It elaborately calibrates the transmission signal according to devices' perceived CSI for heterogeneous power constraints accommodation and CSI attack resilience. We show that by utilizing the common randomness, which deviates from the conventional FL, random sparsification-like dimension reduction can augment privacy in addition to the intrinsic privacy amplification effect of AirComp. We establish near-optimal convergence rates and explicit trade-offs among privacy, communication and utility for PROBE. Finally, extensive experiments on benchmark datasets are conducted to validate our theoretical findings and showcase the superiority of PROBE in realistic settings.

**Index Terms**—Federated learning, over-the-air computing, differential privacy, communication efficiency

## 1 Introduction

With the proliferation of mobile and IoT devices, the network edge is witnessing an unprecedented data explosion. To leverage the potential of the edge big data, it is imperative to implement large-scale machine learning algorithms at the edge [1]–[3], which fosters the concept of edge intelligence and facilitates various innovative applications that improve human welfare, such as smart cities [4], [5], health care [6], [7], and autonomous driving [8], [9]. Federated learning (FL), as a state-of-the-art collaborative learning paradigm, plays a crucial role in achieving edge intelligence and has drawn considerable attention from both academia and industry [10], [11]. In edge FL, the data is distributed across edge devices and processed locally in parallel, which mitigates the risk of direct privacy breach or disclosure. Specifically, a large number of edge devices are connected to an edge server over a shared wireless medium and jointly train a global AI model in successive rounds[1].

However, due to the unreliable nature of wireless connectivity, as well as the constraints in computing and communication resources, the conventional FL scheme that separates communication and computation for aggregation can encounter difficulty in accommodating massive access and incur intolerable latency for many real-time applications. When the number of devices is large, this will cause substantial training efficiency deterioration, which is becoming a hindrance for achieving fast FL at edge. Fortunately, since it only requires to compute the sum/average of the uploads for training the model, over-the-air computing (AirComp) [12] has emerged as a preferable alternative to the standard multi-access communications for edge FL. Unlike the traditional multi-access method that decodes multi-user data independently and performs communication and computation separately, AirComp exploits the waveform superposition property of the multi-access channel to perform computations directly in the air so that fast aggregation of terminal uploads can be achieved. Through the seamless integrated design of communication and computation, AirComp can effectively reduce latency in the distributed train-


- *Y. Tao is with the School of Computer Science and Technology, Shandong University, P.R. China and the School of Electrical Engineering and Computer Science, TU Berlin, Germany.*
  *E-mail: tao@ccs-labs.org*
- *S. Chen, C. Zhang, D. Yu and X. Cheng are with the School of Computer Science and Technology, Shandong University, P.R. China.*
  *E-mail: {szchen, xczhw}@mail.sdu.edu.cn,{dxyu, xzcheng}@sdu.edu.cn*
- *D. Wang is with the Division of CEMSE, King Abdullah University of Science and Technology, Saudi Arabia.*
  *E-mail: di.wang@kaust.edu.sa*
- *F. Dressler is with the School of Electrical Engineering and Computer Science, TU Berlin, Germany.*
  *E-mail: dressler@ccs-labs.org*


*Manuscript received XX; revised XX. Y. Tao was supported in part by the National Science Foundation of China (NSFC) under Grant 623B2068 and the China Scholarship Council (CSC) under Grant 202306220153. D. Yu was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62122042. D. Yu and X. Cheng were supported in part by the Major Basic Research Program of Shandong Provincial Natural Science Foundation under Grant ZR2022ZD02. D. Wang was supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, FCC/1/1976-49-01 from CBRC of King Abdullah University of Science and Technology (KAUST). He was also supported by the funding RGC/3/4816-09-01 of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). F. Dressler was supported in part by the Federal Ministry of Education and Research (BMBF, Germany) within the 6G Research and Innovation Cluster 6G-RIC under Grant 16KISK020K.*

1. In this paper, we use the terms *round* and *iteration* interchangeably to refer to the same concept.

ing process and improve the training efficiency of edge FL. As a result, federated learning based on AirComp, *i.e.*, over-the-air federated learning (OTA-FL), has recently gained considerable interest in edge intelligence research [13]–[18].

### 1.1 Challenges and Prior Art

Despite the benefits of OTA-FL, there are still some serious challenges that need to be tackled. One of the most critical ones is the privacy preservation of the local model updates. Although OTA-FL circumvents direct sharing of sensitive local data, various privacy attacks have been developed to infer private information from these datasets. For instance, membership inference attacks [19], [20] can ascertain whether a specific data sample belongs to a private local training dataset, while reconstruction attacks [21]–[23] can potentially recover the entire private local training dataset. Therefore, it is vital to provide rigorous privacy protection for the shared gradients in OTA-FL.

To address this issue, a widely adopted technique is differential privacy (DP) [24], a well-established technique in machine learning that offers quantifiable privacy guarantees. Several efforts have been made to study OTA-FL under DP constraint. [25] demonstrated that channel noise can offer free privacy for OTA-FL when the DP requirements are mild. If the channel noise is insufficient to meet the desired DP levels, [26] proposed a noisy gradient descent-based algorithm and proved that artificial noises added by each device can enhance the privacy of all devices. This also highlights AirComp's potential for privacy amplification, which is intuitive since the edge server only receives a superimposed gradient without knowing each specific one.

However, in [25], [26], channel state information (CSI) at the devices plays a pivotal role not only in aligning their gradients at the central server but also in the privacy guarantee. In practice, devices obtain CSI from pilots transmitted by the server, which gives rise to a vulnerability, as the server, acting as an adversary, can manipulate the pilots to degrade the privacy level. To cope with this, [27] designed a protocol that is resilient to CSI attacks, but it neglected the specific power constraints of devices and lacked convergence guarantees. Their assumption that only devices with high perceived channel coefficients participate raises concerns, especially in stringent DP cases, where heavy noise injection may surpass affordable power limits, even for devices with strong channels. Moreover, the lack of convergence guarantee also impedes practical implementation of their protocol. Achieving enhanced privacy with both CSI resilience and compatibility to device power limitations thus remains an open challenge.

Another pressing concern for OTA-FL is the limited communication bandwidth at edge. While AirComp enhances channel resource utilization and reduces training latency, edge bandwidth is still scarce. Given the trend towards complex, high-dimensional learning models [28] and emerging large generative AI models at the edge [29], [30], such as ChatGPT [31] and DALL-E [32], transmitting high-dimensional local model updates or gradient from numerous edge devices to the central server becomes increasingly challenging for OTA-FL. Hence, sender devices must strategically craft and compress uploads to fit bandwidth constraints in coordination with the edge server for accurate model learning.

Recent efforts have been made to address this problem in non-private cases. For example, [33] proposed an innovative strategy that involves sparsifying gradient estimates. In their approach, these estimates are projected into a lower-dimensional space aligned with the available channel bandwidth. Then, the server reconstructs the sum of gradient estimates from the aggregated compressed sparse gradients using the principles of compressive sensing. Similarly, [34] presented a band-limited coordinated descent approach that also entails integrating gradient sparsification to mitigate communication costs while optimizing edge-to-server communication effectively.

However, when considering both privacy and bandwidth limitation for OTA-FL, things get complicated. Compression techniques prolong training iterations due to errors, and privacy guarantees typically deteriorate with increased iterations. To ensure a certain target level of privacy, more perturbation must be injected for each round which then makes the convergence error become larger and thus degrades the learning accuracy. The crux of the matter lies in mitigating these conflicts and finding an optimal balance between communication overheads and privacy. Notably, this challenge has received limited attention so far, with [35] being a notable exception. [35] used Johnson-Lindenstrauss (JL) random projection to reduce the dimension of the local gradient estimates. However, their gradient projection process relies on matrix multiplication, incurring high computation costs for edge devices, especially with high-dimensional models. Moreover, their method also depends on the knowledge of true CSI information at edge devices, making it vulnerable to CSI attacks and potentially compromising their privacy claims.

Therefore, it is imperative to consider both privacy and communication band adaptation for OTA-FL at edge, which remains largely unexplored in existing works. Specifically, we aim to bridge this gap by addressing the following question:

*How to achieve private and band-adapted OTA-FL algorithm with lightweight computation, while also offering AirComp-incurred privacy amplification, CSI attack resilience, and nearly optimal learning utility guarantees?*

### 1.2 Main Contributions

In this paper, we provide a positive answer to the above question. We consider a realistic edge scenario where a massive number of edge devices with limited computing and memory resources linear with the model dimension $d$. They cooperatively train a global machine learning model with the assistance of an edge server over a wireless multi-access channel with a scarce band such that only $p$ ($< d$) orthonormal baseband waveforms are available. To make OTA-FL viable in this setting, we devise a novel OTA-FL algorithm that is endowed with a newly developed over-the-air gradients aggregation rule.

Our main contributions can be summarized as follows:

To enable private OTA-FL at band-limited edge, we design a lightweight over-the-air gradients aggregation rule, called `PB-O-GAR`, which is the core

of OTA-FL. PB-O-GAR combines the idea of band-tailored sparsification-like dimension reduction and random Gaussian noise perturbation to provide rigorous device-level differential privacy and band-adapted communication at the same time. Furthermore, by elaborately designing the transmission signal calibration coefficient, PB-O-GAR accommodates the heterogeneous power constraints across edge devices and enables resilience against CSI attacks in AirComp.

Building upon PB-O-GAR, a novel OTA-FL algorithm PROBE is then proposed. Our learning algorithm follows the popular federated stochastic gradient descent framework, which ensures its easy implementation in practice. Moreover, by utilizing the common randomness generated at the server, which deviates from the conventional FL, PROBE makes the random sparsification-like dimension reduction able to augment privacy protection in addition to the intrinsic privacy amplification effect of AirComp.

We establish nearly optimal learning utility bounds and explicit trade-offs among privacy, communication, and utility for both objective functions satisfying Polyak-Łojasiewicz (PL) condition and general non-convex functions. In addition, we perform extensive experiments on benchmark datasets to corroborate our theoretical findings and illustrate the superior performance of PROBE in various settings.

### 1.3 Organization

The remainder of this paper is structured as follows. In Section 2, we provide the necessary background, including the FL setup, the AirComp and communication model, the DP and threat model, and the notation used throughout the paper. In Section 3, we propose the Over-the-air Gradients Aggregation Rule PB-O-GAR and the OTA-FL algorithm PROBE. In Section 4, we present the main theoretical results, including the power compatibility condition, the privacy guarantees and the utility bounds for the proposed algorithms. In Section 5, we conduct extensive experiments to evaluate the performance of PROBE and compare it with the state-of-the-art baseline method. In Section 6, we summarize the paper and discuss future directions.

## 2 PRELIMINARIES

### 2.1 Federated Learning at the Edge

A typical edge FL system consists of $m$ edge devices and a central edge server, collaboratively building a shared machine learning model. Each device $i \in [m]$ has a local dataset $D_i$ that comprises $n$ data samples $\{\xi_{i;1}; \ldots; \xi_{i;n}\}$. Notably, these local datasets cannot be shared during the learning process. The learning objective can be formulated as the following empirical risk minimization problem:

$$\min_{w \in \mathbb{R}^d} L(w) := \frac{1}{m} \sum_{i=1}^{m} L_i(w). \tag{1}$$

The goal is to find an optimal model parameter $w$ in the $d$-dimensional Euclidean space $\mathbb{R}^d$ that minimizes the global empirical risk of the loss on the union of all local datasets. Here $L_i(\cdot)$ denotes the local empirical risk function pertaining to each device $i \in [m]$. More specifically, for any model parameter $w$, $L_i(w)$ can be expressed as

$$L_i(w) := \frac{1}{n} \sum_{j=1}^{n} \ell(w; \xi_{i;j}); \tag{2}$$

where $\ell(w; \xi_{i;j})$ is the *sample-wise* loss function of model parameter $w$ for data point $\xi_{i;j}$. For any data sample $\xi$ and model parameter $w$, we assume that $\ell(\cdot; \cdot)$ is component-wise $L = \sqrt{d}$-bounded gradient, *i.e.*,

$$|\nabla_k \ell(w; \xi)| \le L = \sqrt{d}, \forall k \in [d]. \tag{3}$$

Note that, (3) is fairly standard and has been extensively used in the previous FL literature, *e.g.*, [36], [37].

To solve (1), the classical method is federated stochastic gradient descent (FedSGD) [38]. In FedSGD, the devices iteratively update the model parameter under the orchestration of the central server, resulting in a sequence of $w_0; w_1; \ldots$. In the $t$-th iteration, the central server first broadcasts the current model $w_{t-1}$ to all the devices. Then each device $i$ computes an unbiased stochastic gradient $g_{i;t}$ such that

$$\mathbb{E}[g_{i;t}] = \nabla L_i(w_{t-1}) := \frac{1}{n} \sum_{j=1}^{n} \nabla \ell(w_{t-1}; \xi_{i;j}); \tag{4}$$

and sends $g_{i;t}$ to the central server. Finally, the central server updates the model parameter via the gradient descent step:

$$w_t = w_{t-1} - \eta_t \frac{1}{m} \sum_{i=1}^{m} g_{i;t}; \tag{5}$$

where $\eta_t$ is the learning rate at iteration $t$. The iteration proceeds until some termination condition is reached.

### 2.2 Over-the-Air Computing and Communication Model

We adopt the analog over-the-air computing [12] for the training of federated learning model. The essence of over-the-air computing (AirComp) is to exploit the waveform superposition property of multi access channel, where devices modulate the gradient on the waveform and use the air as an auto aggregator. To illustrate the principle of applying AirComp in FL, we first consider the ideal case with sufficient bandwidth, and describe how devices upload their full gradients to the server at $t$-th iteration in general. Recall that the central server first broadcasts the latest $d$-dimensional global model $w_{t-1}$ to all devices. Owing to the high power available at the central server, we assume the global model is error-free when received by all devices. Then each device $i$ calculates its local gradient $g_{i;t}$ and modulates it onto $d$ orthonormal waveforms, one for each component of the gradient vector. Specifically, the analog signal constructed by device $i$ at time $\tau$ $(0 < \tau < \cdot)$, denoted as $x_i(\tau)$, can be defined as follows,

$$x_i(\tau) := \langle s(\tau); g_{i;t} \rangle; \tag{6}$$

where $\langle \cdot; \cdot \rangle$ denotes the inner product between two vectors and $s(\tau) = (s_1(\tau); s_2(\tau); \ldots; s_d(\tau))$ is a set of orthonormal baseband waveforms that satisfies:

$$\int_0^{\cdot} s_k(\tau) \, d\tau = 1, \text{ for } k = 1; \ldots; d; \tag{7}$$

$$\int_0^{} s_k(\ )s_{k'}(\ )\,\mathrm{d}\ = 0;\ \text{for } k \neq k'. \tag{8}$$

Essentially, the signal $x_i(\ )$ is a superposition of the analog waveforms whereas the magnitude of $s_k(\ )$ equals the $k$-th component of $g_{i;t}$. All the waveforms $\{x_i\}_{i=1}^m$ are sent concurrently from devices into the spectrum. The received signal of the central server at time , denoted by $y(\ )$ can be expressed as follows:

$$y(\ ) := \sum_{i=1}^{m} c_i\ x_i(\ ) + z(\ );\tag{9}$$

where $c_i$ is the time-invariant channel state information (CSI) of device $i$, and $z(\ )$ is channel interference. We assume that the transmitters perfectly know and correct the phase shift in their channels. Therefore we consider real channel coefficients, *i.e.*, $c_i \in \mathbb{R}^+$. The received signal $y(\ )$ at the server will be passed to a set of matched filters, where each of them is tuned as $s_k(\ )$. Denote the vector sent by each device $i$ at each training iteration $t$ by $x_{i;t}$ (which equals $g_{i;t}$ here) and the received vector at the server at iteration $t$ by $y_t$. Then we can formulate the input-output relationship for $t$-th training iteration as follows:

$$y_t = \sum_{i=1}^{m} c_i\ x_{i;t} + z_t = \sum_{i=1}^{m} c_{i;t}\ \ g_{i;t} + z_t;\tag{10}$$

where $z_t \sim \mathcal{N}(0;\ \sigma_0^2)$ is a $d$-dimensional unbiased Gaussian noise due to channel inference. In practice, devices acquire CSI $c_i$'s from pilots transmitted by the central server.

Let $p$ be the number of available orthonormal baseband waveforms, which is determined by the wireless bandwidth. In this paper, we consider a more realistic scenario where communication bandwidth at edge is limited. That is, we assume $p < d$ such that only partial dimensions of the gradients can be transmitted by each device at each iteration. Moreover, each device has limited power per iteration such that each transmitted vector $x_{i;t}$ is subject to an average power constraint $P_i$, *i.e.*,

$$\mathbb{E}[\|x_{i;t}\|_2^2] \le P_i;\ \ \forall i \in [m], t \in [T].\tag{11}$$

Let $\alpha_i := P_i c_i^2$ denote the true effective SNR of device $i$, and let $\hat{\alpha}_i := P_i \hat{c}_i^2$ denote the effective SNR perceived by device $i$ from the pilots transmitted by the server. In practice, each true CSI $c_i$ cannot be infinitely large, thus we assume that there exists a known upper bound $b$ for all the $c_i$'s. Furthermore, let $b := (\max_i P_i)\, b^2$ be the upper bound of the true effective SNR among all the devices, which is a common information in the system.

### 2.3 Differential Privacy and Threat Model

Differential privacy provides provable privacy guarantees and is resilient to arbitrary auxiliary information available to attackers. We denote $D \sim D'$ as a pair of adjacent datasets, which means that $D'$ can be obtained from $D$ by changing only one record.

**Definition 1** ($(\varepsilon;\ \delta)$-DP [39])**.** A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies $(\varepsilon;\ \delta)$-differential privacy if for every pair of adjacent datasets $D \sim D'$ that differ in exactly one data record, it holds for $\forall S \subseteq \mathcal{Y}$ that

$$\mathrm{P}(M(D) \in S) \le e^{\varepsilon}\,\mathrm{P}(M(D') \in S) + \delta.\tag{12}$$

In FL, DP guarantees can be categorized into *instance-level* DP (protecting each instance in the dataset of any device) and *device-level* DP (protecting the whole dataset of any device), depending on how the adjacent dataset is defined. In this paper, we are interested in device-level DP. That is, we want to ensure that the aggregated global gradients are nearly the same regardless of the change of any local dataset at any device. In this case, each element of $D$ or $D'$ is a local dataset. To achieve DP, one can inject appropriate Gaussian noise into the output, which is called Gaussian Mechanism.

**Definition 2** (Gaussian Mechanism [39])**.** Given any input data $D \in \mathcal{X}^n$ and a query function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism $M_G$ is defined as $q(D) + Z$ where $Z \sim \mathcal{N}(0;\ \sigma_G^2 \mathbf{I}_d)$. Let $\Delta_2(q)$ be the $\ell_2$-sensitivity of $q$, *i.e.*, $\Delta_2(q) := \sup_{D \sim D'} \|q(D) - q(D')\|_2$. For any $\varepsilon;\ \delta > 0$, $M_G$ guarantees $(\frac{\Delta_2(q)}{\sigma_G}\sqrt{2\log\frac{1.25}{\delta}};\ \delta)$-DP. That is, if we want the output of $q$ to be $(\varepsilon;\ \delta)$-DP for any $0 < \varepsilon < 1$, then $\sigma_G$ should be set to $\frac{\Delta_2(q)}{\varepsilon}\sqrt{2\log\frac{1.25}{\delta}}$.

In the context of FL, if $(\varepsilon;\ \delta)$-DP is guaranteed for each training iteration, then the total privacy guarantee over $T$ iterations can be given by the adaptive composition result.

**Lemma 1** (Adaptive Composition [39])**.** For any $\varepsilon_0;\ \delta_0;\ \delta' > 0$, let $M_T = (M_1;\ \ldots;\ M_T)$ be a sequence of $(\varepsilon_0;\ \delta_0)$-DP algorithms, where $M_i$'s are potentially chosen sequentially and adaptively. Then $M_T$ is $(\varepsilon;\ \delta)$-DP, where $\varepsilon = \varepsilon_0\sqrt{2T\log(1/\delta')} + T\varepsilon_0\frac{e^{\varepsilon_0}-1}{e^{\varepsilon_0}+1}$ and $\delta = T\delta_0 + \delta'$. That is, if we want $M_T$ to be $(\varepsilon;\ \delta)$-DP, then it suffices for each $M_i$ to be $(\frac{\varepsilon}{2\sqrt{2T\log(2/\delta)}};\ \frac{\delta}{2T})$-DP.

We assume the central server to be *semi-honest* or *honest-but-curious*. That is, the server will follow the prescribed protocol strictly, but try to infer sensitive information about single individual data samples from both the transmitted gradients and any other available auxiliary information during the FL process. In practice, devices acquire CSI from pilots transmitted by the central server. Therefore, the central server can attack the CSI estimation process to encourage the devices to increase their transmit power or to add less noise to their transmissions by suggesting that their channel quality is worse than it is in reality. We assume that a common pilot signal is used to learn the CSI by all the devices and thus the CSI value can only be scaled by the same parameter across the devices. Let $\hat{c}_i$ be the CSI learnt by device $i$, then $\hat{c}_i = \rho c_i$ for all $i \in [m]$, where $\rho \in (0;1]$ is the scaling parameter.

### 2.4 Other Useful Notations

For any vector $x$, we use $[x]_k$ to denote the $k$-th component of $x$, if it makes sense. Moreover, we use $[x]_k$ to denote the $k$-th non-zero component of $x$, if it makes sense. We use $\odot$ to denote element-wise multiplication for two vectors with equal length, *e.g.*, for any two vectors $x = ([x]_1;[x]_2;\ \ldots;[x]_d)$ and $y = ([y]_1;[y]_2;\ \ldots;[y]_d)$ in $\mathbb{R}^d$, we have $x \odot y = ([x]_1\ [y]_1;[x]_2\ [y]_2;\ \ldots;[x]_d\ [y]_d)$. For any positive integer $N$, we use $[N]$ to denote the set of $\{1;2;\ \ldots;N\}$. For any set $A$, we use $|A|$ to denote its

TABLE 1: Summary of the main notations

| Symbol | Description |
|---|---|
| $m$ | Number of devices |
| $d$ | Dimension of the model |
| $\mathcal{L}(w)$ | Global empirical risk function w.r.t model $w$ |
| $\mathcal{L}_i(w)$ | Local empirical risk function of device $i$ w.r.t model $w$ |
| $\ell(w;\ )$ | Sample-wise loss function w.r.t model $w$ under data |
| $w_t$ | Updated model parameter in round $t$ |
| $g_{i;t}$ | Stochastic gradient generated by device $i$ in round $t$ |
| $g^0_{i;t}$ | Compressed version of $g_{i;t}$ for band adaptation |
| $\tilde{g}_{i;t}$ | Noisy version of $g^0_{i;t}$ for privacy |
| $\tilde{g}_{i;t}$ | Re-scaled version of $\tilde{g}_{i;t}$ |
| $h_i$ | Signal calibration coefficient of device $i$ |
| $x_{i;t}$ | Vector sent by device $i$ in round $t$ |
| $y_t$ | Vector received by the server in round $t$ |
| $_t$ | Learning rate in round $t$ |
| | Scaling parameter of CSI attack |
| $P_i$ | Power limit of device $i$ |
| $c_i$ | True CSI of device $i$ |
| $\mathbf{e}_i$ | CSI learnt by device $i$ |
| $\mathbf{b}$ | Known upper bound for true CSI |
| $_i$ | True effective SNR of device $i$ |
| $\mathbf{e}_i$ | Effective SNR perceived by device $i$ |
| | A Lower bound for perceived effective SNR $\mathbf{e}_i$'s |
| $^0$ | $= {}^2$ |
| $\mathbf{b}$ | Known upper bound of the true effective SNR $_i$'s |
| $;$ | Differential privacy parameters |
| $p$ | Number of available orthonormal baseband waveforms |
| | Compression ratio $p=d$ |
| $z_t$ | Channel noise in round $t$ |
| | Standard deviation of Gaussian noise added by each device |

cardinality. Table 1 summarizes the main notations used in this paper.

**Algorithm 1:** PB-O-GAR: Private and Band-Adapted Over-the-Air Gradients Aggregation Rule (for $t$-th iteration)

**Input:** Local gradients $\{g_{i;t}\}_{i=1}^m$, indicator vector $I_C$ for the component-index set $C$, noise magnitude .

$p \quad \|I_C\|_1$;
**for every** *device* i **in parallel do**
  /* **Band-Adapted Dimension Reduction** */
  **for** $k \quad 1; \quad ; p$ **do**
    $[g^0_{i;t}]_k \quad [g_{i;t}]_{[I_C]_k}$;
  /* **Gaussian Perturbation for Privacy** */
  Sample $v_i \quad N(0; \ ^2\mathbf{I}_p)$;
  $\tilde{g}_{i;t} \quad g^0_{i;t} + \ _{i;t}$;
  /* **Gradient Re-Scale and Signal Calibration** */
  Calculate the compression ratio $p=d$;
  Get the re-scaled gradient $\tilde{\mathbf{g}}_{i;t} \quad \frac{1}{} \tilde{g}_{i;t}$;
  Set the vector to send as $x_{i;t} \quad h_i \ \tilde{\mathbf{g}}_{i;t}$ with $h_i = \frac{1}{\mathbf{e}_i}\sqrt{\frac{L^2+d\ ^2}{}}$;
  Send $x_{i;t}$ to the server simultaneously with other devices via AirComp;
/* **Post-Processing at Server** */
**Server** receives the vector $y_t$ from the channel;
**Server** initializes $\tilde{\mathbf{g}}_t \quad 0$;
**for** $k \quad 1; \quad ; p$ **do**
  **Server** sets $[\tilde{\mathbf{g}}_t]_{[I_{C_t}]_k} \quad \frac{1}{m} [y_t]_k; 8k \ 2 \ [p]$;

**Output:** The global gradient estimator $\tilde{\mathbf{g}}_t$.

## 3 Our Proposed Methods

### 3.1 Over-the-Air Gradients Aggregation Rule

To address privacy and communication aspects simultaneously for local gradients sharing, we need to answer: (1) how to release a private and bandwidth adaptive compressed local gradient at each device and (2) how to get a reasonable estimator for the global gradient at the server. Specifically, given a model parameter $w \ 2 \ \mathbb{R}^d$, how can devices privately release local gradient estimators in a lower-dimensional space $\mathbb{R}^p$ with $p < d$ such that these local gradient estimations, when uploaded by AirComp, can be received at the server as an unbiased global gradient estimation? We solve this by presenting a private and band-adapted over-the-air gradients aggregation rule called PB-O-GAR, which will lie in the center of our proposed learning algorithm.

We present PB-O-GAR in Algorithm 1. Specifically, PB-O-GAR requires a component-index set $C \quad \{1; \quad ; d\}$ as input, among others, indicating which components are active after the band-adapted compression. The set $C$ can be fully described by an indicator vector $I_C \ 2 \ \{0;1\}^d$, which is defined as follows:

$$[I_C]_k = \begin{cases} 1 & \text{if } k \ 2 \ C \\ 0; & \text{otherwise} \end{cases}; 8k \ 2 \ [d]: \tag{13}$$

By using $I_C$, the representation, communication and storage of $C$ will cost only $O(d)$ bits, which is lightweight. The gradient sharing and aggregation process in PB-O-GAR can be divided into five steps, which are elaborated as follows.

**Step 1. Band-Adapted Dimension Reduction:**

Given the components-index set $C$ (represented by $I_C$), each device $i$ reduces the dimension of local gradient $g_{i;t}$ by only keeping the active components indicated by $I_C$ to get a $p$-dimensional vector $g^0_{i;t}$:

$$[g^0_{i;t}]_k = [g_{i;t}]_{[I_C]_k}; 8k \ 2 \ [p]: \tag{14}$$

**Step 2. Gaussian Perturbation for Privacy:**

To protect the privacy of local data samples, each device $i$ perturbs the compressed local gradient with a Gaussian random noise $_{i;t}$ of magnitude :

$$\tilde{g}_{i;t} = g^0_{i;t} + \ _{i;t} \text{ with } \ _{i;t} \quad N(0; \ ^2\mathbf{I}_p): \tag{15}$$

**Step 3. Gradient Re-Scale and Signal Calibration:**

Each device $i$ further magnifies $\tilde{g}_{i;t}$ with the factor of $1=$ to get an unbiased estimator $\tilde{\mathbf{g}}_{i;t}$ for its local gradient $g_{i;t}$, where $:= p=d$ is referred to as the compression ratio:

$$\tilde{\mathbf{g}}_{i;t} = \frac{1}{} \tilde{g}_{i;t}: \tag{16}$$

Based on this estimator, each device $i$ then generates the vector to send, *i.e.*, $x_{i;t}$, as

$$x_{i;t} = h_i \ \tilde{\mathbf{g}}_{i;t}; \tag{17}$$

where the scaling factor $h_i := \frac{1}{\mathbf{e}_i}\sqrt{\frac{L^2+d\ ^2}{}}$ is the signal calibration coefficient that serves for fitting power constraints across devices as well as aligning gradients at the server. After that, all devices modulate $x_{i;t}$'s onto $d$ orthonormal waveforms, one for each component of $x_{i;t}$ and transmit their analog signals simultaneously.

**Step 4. Post-Processing at Server:**

After receiving the superposition of the signals transmitted by the devices, the server gets a vector $y_t$ as follows:

$$y_t = \sum_{i=1}^{m} c_i \, x_{i;t} + z_t = \sum_{i=1}^{m} c_i \, h_i \, g_{i;t} + z_t \tag{18}$$

$$= \sum_{i=1}^{m} \frac{c_i}{e_i} \sqrt{\frac{}{L^2 + d^2}} \, g_{i;t} + z_t \tag{19}$$

$$= \frac{1}{} \sqrt{\frac{}{L^2 + d^2}} \sum_{i=1}^{m} g_{i;t} + z_t; \tag{20}$$

where $z_t \sim N(0; \sigma_0^2 I_d)$ is the channel noise. For simplicity, we define

$$:= \frac{1}{} \sqrt{\frac{}{L^2 + d^2}} = \sqrt{\frac{}{L^2 + d^2}}; \tag{21}$$

then the received vector $y_t$ can be further expressed as

$$y_t = - \sum_{i=1}^{m} g_{i;t}^{0} + - \sum_{i=1}^{m} {}_{i;t} + z_t = - \sum_{i=1}^{m} g_{i;t}^{0} + e_t; \tag{22}$$

where $e_t \sim N(0; e^2)$ is the effective noise in total with e being $\sqrt{\frac{2m}{2} {}^2 + \sigma_0^2}$. The server performs post-processing on $y_t$ to recover a $d$-dimensional unbiased global gradient estimator $g_t \in \mathbb{R}^d$ for the global gradient $r L(w_{t-1})$. Specifically, the server first initializes $g_t = 0$, and then assigns normalized value of each component of $y_t$ to the corresponding position of $g_t$ as follows:

$$[g_t]_{[I_{C_t}]_k} = \frac{1}{m} [y_t]_k; \; 8k \in [p]: \tag{23}$$

**Remark 1** (Sparsification-like dimension reduction)**.** We rethink the global gradient estimate in a more concise form. Let $v_{i;t}^{+} \sim N(0; {}^2 I_d)$ and $z_t^{+} \sim N(0; \sigma_0^2 I_d)$ be the complement of $v_{i;t}$ and $z_t$ in the original $d$-dimensional space, *i.e.*,

$$[v_{i;t}^{+}]_{[I_{C_t}]_k} = \frac{1}{m} [v_{i;t}]_k; \; 8k \in [p]; \tag{24}$$

$$[z_t^{+}]_{[I_{C_t}]_k} = \frac{1}{m} [z_t]_k; \; 8k \in [p]: \tag{25}$$

The global gradient estimate $g_t$ can be written as

$$g_t = \frac{1}{} \left( \frac{1}{m} \sum_{i=1}^{m} (g_{i;t} + {}_{i;t}^{+}) \right) I_{C_t} + \frac{z_t^{+}}{m} I_{C_t}; \tag{26}$$

which indicates that the global gradient estimate $g_t$ essentially comes from a *sparsification-like* compression together with Gaussian perturbation.

### 3.2 OTA-FL Algorithm

With the over-the-air gradients aggregation rule PB-O-GAR, we now present our learning algorithm PROBE for private OTA-FL at Band-Limited Edge in Algorithm 2. Our learning algorithm consists of two phases:

**Phase 1. Initialization:**

As explained in Section 2.3, the CSI attack causes each device to perceive a distorted version of effective SNR. Let be a common lower bound for all the perceived effective SNRs $e_i$'s such that $e_i \; ; 8i \in [m]$. To obtain , we design a two-pass communication scheme between the devices and the server. In the first pass, all devices send their perceived effective SNR, *i.e.*, $e_i$ to the server. Then in the second pass, the server broadcasts the common lower bound to all devices. The validity of the received can be easily verified by each device by checking if is less than or equal to its perceived effective SNR $e_i$. Furthermore, we define $_0 := = {}^2$, which is then a lower bound for all true effective SNR, *i.e.*, $_0 \; _i$. We note that, $_0$ is bounded above by the publicly known maximal true effective SNR b.

**Algorithm 2:** PROBE: Private OTA-FL at Band-Limited Edge

**Input:** Local datasets $\{D_i\}_{i=1}^m$, initial model $w_0$, noise magnitude , number of available orthonormal baseband waveforms $p$, number of learning iterations $T$, learning rate $_t$.

/* Initialization */

**for every** *device* i **in parallel do**
  Perceive $e_i$;
  Send $e_i$ to the server;

**Server** broadcasts a uniform lower bound for all the received $e_i$'s;

/* Learning Process */

**for iteration** $t \; 1; \; ; T$ **do**
  **Server** generates a random indicator vector $I_{C_t} \in \{0,1\}^d$ such that $\|I_{C_t}\|_1 = p$;
  **Server** broadcasts the latest model $w_{t-1}$ and indicator vector $I_{C_t}$;
  **for every** *device* i **in parallel do**
    Compute local gradient $g_{i;t} \; r L(w_{t-1})$;
  PB-O-GAR$(\{g_{i;t}\}_{i=1}^m; I_{C_t}; )$;
  **Server** updates the model via $w_t \; w_{t-1} \; _t \, g_t$;

**Output:** $w_{\text{output}}$

**Phase 2. Model Training**

The model training process consists of a series of synchronized iterations $t = 1; \; ; T$. In each iteration $t \in [T]$, the central server randomly generates a component-index set $C_t \; [d]$ whose size accommodates the band limit such that $|C_t| = d \; p$, where is referred to as the compression ratio. The server broadcasts the indicator vector $I_{C_t}$ together with the latest model $w_{t-1}$ to all the devices. The devices compute local gradients, which are then shared to and aggregated at the server via our PB-O-GAR. Finally, the server updates the global model via the gradient descent step of $w_t = w_{t-1} \; _t \, g_t$, where $_t$ is the learning rate.

## 4 Main Results

### 4.1 Power compatibility

The power of edge devices is limited and their limitations are often diverse among devices, which makes the implementation of OTA-FL subject to these power limitations. So it is essential for the OTA-FL algorithm design to account for the heterogeneous power constraints. Now, we demonstrate that our PROBE satisfies the heterogeneous power constraints among devices, from which we also disclose how PROBE achieves the power compatibility via our novel

signal scaling rule. Moreover, we also compare our approach for achieving power compatibility with the previous works on private OTA-FL in terms of privacy preservation.

To respect the power limitation at each device, we let each device $i$ scale its private and compressed local gradient $g_{i;t}$ by the coefficient $h_i$, which depends on both the public effective SNR information and the local effective SNR information $e_i$, among others. This way, the transmitted local gradient estimates can be properly aligned so that each local gradient has equal weight in the received vector at the server as shown in (22), which is essential for the server to obtain an unbiased global gradient estimate.

**Theorem 1** (Power Compatibility)**.** Our algorithm satisfies the power constraints of devices, *i.e.*, for each $i \in [m]$ and $t \in [T]$, $\mathsf{E}[\|x_{i;t}\|_2^2] \le P_i$.

*Proof of Theorem 1.*

$$\mathsf{E}[\|x_{i;t}\|_2^2] = \mathsf{E}\left\| h_i \left( \frac{1}{L} g_{i;t} \right) \right\|_2^2 \tag{27}$$

$$= \frac{h_i^2}{L^2} \mathsf{E}\left\| g_{i;t} \right\|_2^2 \tag{28}$$

$$= \frac{h_i^2}{L^2} \mathsf{E}\left\| g^0_{i;t} + \Delta_{i;t} \right\|_2^2 \tag{29}$$

$$\le \frac{h_i^2}{L^2} \left( d \frac{L}{\sqrt{d}} \right)^2 + d \sigma^2 \tag{30}$$

$$= \frac{h_i^2}{e_i^2} \frac{P_i e_i^2}{e_i^2} = P_i. \tag{31}$$

□

**Remark 2** (Comparison with existing designs)**.** We note that, to satisfy the power limitation and meanwhile get an unbiased global estimate at the server, another strategy is used in [26] as well as its follow-up work [35]. Specifically, they let the vector to send at each device be

$$x_{i;t} = \frac{\sqrt{\alpha_{i;1} P_i}}{L} g_{i;t} + \sqrt{\alpha_{i;2} P_i}\, \Delta_{i;t}. \tag{32}$$

Here $\alpha_{i;1} \in [0, 1]$ and $\alpha_{i;2} \in [0, 1 - \alpha_{i;1}]$ denote the fraction of power dedicated to the normalized local gradient vector $\frac{1}{L} g_{i;t}$ and the DP noise $\Delta_{i;t} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, respectively. These parameters satisfy $\alpha_{i;1} + \alpha_{i;2} \le 1$ so that the maximum power constraint of $P_i$ is satisfied. In order to form an unbiased estimate of the global gradient, all devices pick the $\alpha_{i;1}$'s as:

$$\alpha_{i;1} = \frac{\min_{i' \in [m]} h_{i'}^2}{h_i^2}; \tag{33}$$

so that the server can receive an aggregated vector where all local gradients have the same weight:

$$y_t = \frac{\sqrt{\min_{i' \in [m]} h_{i'}^2}}{L} \sum_{i=1}^{m} g_{i;t} + \sum_{i=1}^{m} \sqrt{\alpha_{i;2} P_i}\, \Delta_{i;t} + z_t. \tag{34}$$

In their design, the level of privacy of each device is determined by the power dedicated to the DP noise, which is constrained by the ratio of its individual effective SNR and the worst effective SNR across devices. When the power constraints across devices become more consistent, the privacy protection provided by DP noises becomes weaker. In the worst-case scenario of homogeneous effective SNR such that $\alpha_{i;1} = 1$ for all devices, no power is left for privacy preservation. To overcome this drawback, we do not divide power between gradient and noise, but treat the gradient and the noise as a whole by first generating the noisy gradient and then re-scale the noisy gradient according to both global and local effective SNR information. By the expertly designed coefficients $h_i$'s, we achieve the power compatibility and unbiased global estimate simultaneously.

### 4.2 Privacy Guarantees

For the privacy part, we start by presenting per-device privacy loss guaranteed in each iteration by injecting DP noise of magnitude at each device.

**Theorem 2** (Per-Round Individual Privacy Loss)**.** For any $\hat{\varepsilon} > 0$, our learning algorithm PROBE achieves $(\varepsilon_t, \hat{\varepsilon})$-LDP for each iteration, where

$$\varepsilon_t = \frac{2\sqrt{2L}\sqrt{\log\frac{1.25}{\hat{\delta}}}}{\sqrt{m^2 + \frac{L^2}{b^2}\sigma_0^2 + \frac{d}{b^2}\sigma_0^2}}. \tag{35}$$

*Proof of Theorem 2.* In order to bound the privacy loss, we first analyse the $\ell_2$-sensitivity of the information queried by the server, *i.e.*, $\mathsf{E}[y_t] = \sum_{i=1}^{m} g^0_{i;t}$, which is denoted as $\Delta_2(q)$. To bound $\Delta_2(q)$, we consider any two different local datasets $D_i$ and $D_i^0$ at device $i$, while fixing the datasets (and thus the original local gradients) of the remaining $(m-1)$ devices. $\Delta_2(q)$ can then be bounded as

$$\Delta_2(q) = \max_{D_i, D_i^0} \| q_{jD_i} - q_{jD_i^0} \|_2 \tag{36}$$

$$= \max_{D_i, D_i^0} \| g^0_{i;t}(D_i) - g^0_{i;t}(D_i^0) \|_2 \tag{37}$$

$$\le \max_{D_i, D_i^0} \| g^0_{i;t}(D_i) \|_2 + \| g^0_{i;t}(D_i^0) \|_2 \tag{38}$$

$$\le 2(\sqrt{p} - L) = 2\sqrt{p}L. \tag{39}$$

Note that the total Gaussian noise injected is of the magnitude of $e = \sqrt{\frac{2}{2}m^2 + \sigma_0^2}$. By using Gaussian mechanism provided in Definition 2, we obtain

$$\varepsilon_t = \frac{2\sqrt{2L}\sqrt{\log\frac{1.25}{\hat{\delta}}}}{\sqrt{m^2 + \frac{L^2}{\rho_0}\sigma_0^2 + \frac{d}{\rho_0}\sigma_0^2}}; \tag{40}$$

which then leads to (35) due to $\rho_0 \ge b$. □

Next, we show how to set proper for any given target level of privacy determined by the DP parameters and .

**Theorem 3** (Required Noise Magnitude for Target Privacy)**.** Given target $\varepsilon > 0$, to guarantee per-device $(\varepsilon, \delta)$-DP through $T$ learning iterations, it suffices to set

$$= \frac{8L}{m\varepsilon} \sqrt{\frac{T \log\frac{2.5T}{\delta} \log\frac{2}{\delta}}{m^2 + \frac{d}{b^2}\sigma_0^2}}. \tag{41}$$

*Proof of Theorem 3.* According to the adaptive composition result of DP (Lemma 1), it suffices to ensure $(\varepsilon_0, \delta_0)$-DP in each iteration for each device $i$, where $\varepsilon_0 = \frac{\varepsilon}{2\sqrt{2T \ln(2/\delta)}}$ and $\delta_0 = \frac{\delta}{2T}$. Using the upper bound on privacy loss incurred

by DP noise with magnitude of shown in (35), it suffices to set such that the following hold,

$$2\sqrt{\log\frac{2.5T}{}}\,\Big/\,\sqrt{m^2+\frac{L^2}{b}\frac{2}{0}+\frac{d}{b}\frac{2}{0}\frac{2}{2}}\qquad 0; \tag{42}$$

which gives

$$\sqrt{\frac{64L^2T\log\frac{2.5T}{}\log\frac{2}{2}}{m+\frac{d}{b}\frac{2}{0}}\cdot\frac{L^2}{b}\frac{2}{0}}. \tag{43}$$

□

**Remark 3** (Genuine Robustness againt CSI attacks)**.** In previous works (*e.g.*, [26], [35]), both the gradient aggregation scheme and the privacy guarantee rely on the (true) CSI at devices, which makes their frameworks vulnerable since CSI values are prone to attacks by the adversarial central server. By contrast, in our design, the accurate CSI at devices is not crucial for aligning their model updates at the central server. Also, we have shown in Theorem 2 that the privacy loss is irrelevant to the CSI attack parameter , which means our algorithm is immune to the CSI attack. Recently, [27] also tackles CSI attacks. Our algorithm outweighs theirs in the follwoing three major aspects.

Firstly, to avoid the privacy breach caused by a potential CSI attack, they simply ignore the channel noise, even if the magnitude of channel noise is known. As a result, they require each device inject more noise so that the desired privacy level ( , ) is satisfied at the expense of learning utility. In contrast, we leverage the global information that can be easily obtained via our newly designed two-pass communication process in the initialization phase of PROBE. Based on this, we devise a novel signal rescaling factor $h_i = \frac{1}{e_i}\sqrt{\frac{}{L^2+d^2}}$ instead of directly using the naive $h_i = \frac{1}{e_i}$ as in paper [27]. This way, we reduce the effect of CSI attack in $y_t$, as the weight of the gradient information transmitted by devices, *i.e.*, the factor $-$, is bounded above, which means the adversary cannot make the channel noise insignificantly small in PROBE. Secondly, their method fails to adapt the magnitude of transmit signal at devices to their specific power constraints. In general, to align the local gradients at the server and obtain an unbiased estimate of the global gradient, devices should adjust the magnitude of their transmit signal so that both the desired weight of their local gradients [2] and their heterogeneous power constraints are satisfied. This adjustment is essential for power compatibility and has been used by many previous OTA-FL works, such as [25], [26], [33]–[35]. However, in paper [27], they circumvent this by assuming that devices participate in the learning process with some random probability such that only the devices with good enough channel coefficients can participate in each round to improve power efficiency. And for those participated devices, they have sufficient power, irrespective of their channel states, to transmit the vector $x_{i;t} = \frac{1}{e_i}(g_{i;t} + \ _{i;t})$. This raises serious concerns about the practicality of their method. Due to the CSI attack, $e_i$ could be very small, which makes $kx_{i;t}k_2$ very large and thus requires a lot of power to send $x_{i;t}$. This problem will be aggravated if we consider strict DP cases where heavy noise injection is needed, leading to $k\ _{i;t}k_2$ as well as $kx_{i;t}k_2$ exceeding the power the devices can afford to send and thus resulting in the case where only few or even none of the devices can participate in the learning process. Therefore, their assumption is simple but evidently unrealistic and fatal.

Finally, [27] does not offer any result for learning utility of their algorithm, in both theoretical and experimental aspects. So it remains unclear how effective their OTA-FL algorithm can be even under the strong client participation assumption mentioned above. We provide both theoretical and experimental utility results for PROBE.

Altogether, our work provides the first solution that provides genuine robustness to CSI attacks for private OTA-FL.

**Remark 4** (Privacy amplification)**.** The per-device privacy loss shown in (35) scales as at most $\mathcal{O}(\sqrt{\ }\ ^{1})$. Previous work [26] has shown that the per-device privacy loss can be reduced from $\mathcal{O}(^{1})$ to $\mathcal{O}(\frac{p}{m}\ ^{1})$ by utilizing analog AirComp. In this work, we further reduce the privacy loss by a multiplicative factor $O(^{p})$ via the sparsification-like gradient compression, which enhances the privacy for devices. In [36], [40], sparsification is used in FL with traditional digital aggregation scheme for privacy. The set of active components therein is determined separately by each device, which is reasonable in the digital scheme since model updates are transmitted individually. However, if we simply use their approach in the analog scheme, the amplification brought by AirComp, *i.e.* the $=\sqrt{m}$ factor, will not be preserved. To fix this, we adopt a different approach where the active components are randomly sampled by the central server. Thanks to the common active components across devices, the privacy amplification brought by analog AirComp and the sparsification are retained simultaneously in OTA-FL.

### 4.3 Utility Analysis

#### 4.3.1 Accuracy of the Global Gradient Estimate

As a preparation for studying the utility of PROBE, we study the accuracy guarantee for the global gradient estimate obtained at the server. Specifically, we show that the global estimate $b_t$ used at the server per training iteration is unbiased and has $kg_tk_2$-dependent variance.

**Lemma 2.** For any $t \in [T]$, $b_t$ satisfies that

1) unbiased expectation:

$$\mathsf{E}[b_t] = g_t := \ r\,L(w_{t\ 1}) = \frac{1}{m}\sum_{i=1}^{N} g_{i;t}; \tag{44}$$

2) $kg_tk_2$-dependent variance:

$$\mathsf{E}\ kb_t \quad g_tk_2^2 \qquad \frac{1}{\ }kg_tk_2^2 + \frac{d}{m}\frac{2}{} + \frac{d}{}\frac{2}{0}\Big/ m^2. \tag{45}$$

2. In this paper, we assume all local gradients have equal weight, *i.e.*, the global gradient is the simple average of all the local gradients. However, this can be easily extended to weighted average case by setting the scaling factor $h_i = \frac{m\,q_i}{e_i}\sqrt{\frac{}{L^2+d^2}}$, where $q_i$ is the weight of device $i$.

*Proof of Lemma 2.* We start with the unbiasedness of $\tilde{g}_t$. Specifically, for each component $k \in [d]$, we have

$$\begin{aligned}
\mathrm{E}[[\tilde{g}_t]_k] &= \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}([g_{i,t}]_k + [\xi_{i,t}^+]_k)\,[I_{C_t}]_k + \frac{[z_t^+]_k}{m}\,[I_{C_t}]_k\right] \\
&= \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}([g_{i,t}]_k + [\xi_{i,t}^+]_k)\right] + \mathrm{E}\left[\frac{[z_t^+]_k}{m}\right]\mathrm{E}[[I_{C_t}]_k] = \frac{1}{m}\sum_{i=1}^{m}[g_{i,t}]_k \quad (\because 1 + (1-\beta)\cdot 0) \\
&= \frac{1}{m}\sum_{i=1}^{m}[g_{i,t}]_k.
\end{aligned}$$

Next, we bound the variance of $\tilde{g}_t$:

$$\begin{aligned}
&\mathrm{E}\,\|\tilde{g}_t - g_t\|_2^2 \\
&= \mathrm{E}\left[\left\|\frac{1}{\beta}\,g_t + \frac{1}{m}\sum_{i=1}^{m}\xi_{i,t}^+\,I_{C_t} - g_t + \frac{z_t^+}{m}\,I_{C_t}\right\|_2^2\right] \\
&= \mathrm{E}\left[\left\|\frac{1}{\beta}\,g_t\,I_{C_t} - g_t\right\|_2^2\right] + \mathrm{E}\left[\left\|\frac{1}{\beta}\,\frac{1}{m}\sum_{i=1}^{m}\xi_{i,t}^+\,I_{C_t}\right\|_2^2\right] + \mathrm{E}\left[\left\|\frac{z_t^+}{m}\,I_{C_t}\right\|_2^2\right] \\
&= \mathrm{E}_{C_t}\left[\left\|\frac{1}{\beta}\,(g_t\,I_{C_t}) - g_t\right\|_2^2\right] + \mathrm{E}_{\xi_{i,t}^+}\mathrm{E}_{C_t}\left[\left\|\frac{1}{\beta}\,\frac{1}{m}\sum_{i=1}^{m}(g_{i,t} + \xi_{i,t}^+)\,I_{C_t}\right\|_2^2\right] + \mathrm{E}_{z_t^+}\mathrm{E}_{C_t}\left[\left\|\frac{z_t^+}{m}\,I_{C_t}\right\|_2^2 z_t^+\right] \\
&= \frac{1}{\beta} - 1\;\|g_t\|_2^2 + \mathrm{E}\left[\sum_{k=1}^{d}\left(\frac{1}{\beta}\,\frac{1}{m}\sum_{i=1}^{m}[\xi_{i,t}^+]_k\right)^2\right] + \mathrm{E}\left[\sum_{k=1}^{d}\frac{[z_t^+]_k^2}{\beta^2 m^2}\right] \\
&= \frac{1}{\beta} - 1\;\|g_t\|_2^2 + \frac{1}{\beta}\,\mathrm{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\xi_{i,t}^+\right\|_2^2\right] + \frac{1}{\beta^2 m^2}\mathrm{E}[\|z_t^+\|_2^2] \\
&\le \frac{1}{\beta} - 1\;\|g_t\|_2^2 + \frac{d}{m} + \frac{d}{\beta^2 m^2}.
\end{aligned}$$

□

**Remark 5** (Computation efficiency)**.** Previously, the work [35] proposed to compress private local gradients into a low-dimensional space through random projection matrices of JL transformation type. However, they only obtained the upper bound for the 2nd-order raw moment of the global gradient estimate instead of variance. Furthermore, their approach unavoidably incurs an additional $O(d^3)$ local computation cost for each device, which is not feasible in resource-constrained edge scenarios. In contrast, our sparsification-like compressor requires only $O(d)$ computation for each device, which is more computation efficient.

### 4.3.2 Utility Bounds

We analyse the utility of our algorithm under $(\varepsilon, \delta)$-DP for non-convex but smooth objective functions.

**Definition 3** ($L$-smooth function)**.** We say a differentiable function $f$ is $L$-smooth ($L > 0$) over space $\mathcal{X}$, if for any pair of $x, x' \in \mathcal{X}$, it holds that

$$\|\nabla f(x) - \nabla f(x')\|_2 \le L\|x - x'\|_2. \tag{46}$$

**CASE 1: Functions with Polyak-Łojasiewicz (PL) condition**

In this part, we consider the case where the global empirical risk function satisfies Polyak-Łojasiewicz (PL) condition.

**Definition 4** (Polyak-Łojasiewicz (PL) condition)**.** For function $f$ over space $\mathcal{X}$, suppose $\mathcal{X}^* := \arg\min_{x\in\mathcal{X}} f(x) \neq \emptyset$; and denote $f^* := \min_{x\in\mathcal{X}} f(x)$, then we say $f$ satisfies the Polyak-Łojasiewicz condition if the following holds for some $\mu > 0$,

$$\|\nabla f(x)\|_2^2 \ge 2\mu\,(f(x) - f^*). \tag{47}$$

For smooth $L$ satisfying the PL condition, we use the expected difference between the empirical risk of the algorithm output $w_{\text{output}}$ and the optimal empirical risk value as the utility measurement of the algorithm, which is called expected *excess* empirical risk, *i.e.*, $\mathrm{E}[L(w_{\text{output}}) - \min_w L(w)]$.

**Theorem 4.** [Utility upper bound under PL condition] Suppose $L$ is $L$-smooth and satisfies Polyak-Łojasiewicz condition over $\mathbb{R}^d$. In Algorithm 2, let $\beta$ be as (41), $\eta_t = \eta$, $T > \frac{\beta^2\varepsilon^2}{128b\log\frac{2}{\delta}}$ and $T = O\left(\log\frac{m^2\varepsilon^2}{L^2 d\log\frac{1}{\delta}}\right) = \log\frac{\cdot}{\cdot}$, then we have the following utility bound for $w_{\text{output}} = w_T$.

$$\mathrm{E}[L(w_T) - L(w^*)] = \tilde{\Theta}\left(\frac{d\log^2 m\log\frac{1}{\delta}}{m^2\varepsilon^2}\right); \tag{48}$$

where the $\tilde{\Theta}(\cdot)$ notation omits poly-logarithmic terms and constants.

*Proof of Theorem 4.* By the $L$-smoothness of $L$, we have

$$L(w_t) \le L(w_{t-1}) + \langle g_t, w_t - w_{t-1}\rangle + \frac{L}{2}\|w_t - w_{t-1}\|_2^2.$$

According to the gradient descent step, we have

$$\begin{aligned}
&\mathbb{E}[L(w_t) \quad L(w_{t\ 1})] \\
&\quad {}_t\mathbb{E}[\hbar g_t; \mathbb{b}_t i] + \frac{{}^2_t}{2}\mathbb{E}[k\mathbb{b}_t k_2^2] \\
&= \quad {}_t\mathbb{E}[kg_t k_2^2] + \frac{{}^2_t}{2}\mathbb{E}[k\mathbb{b}_t \quad g_t k_2^2] + \frac{{}^2_t}{2}\mathbb{E}[kg_t k_2^2] \\
&= \quad \frac{{}^2_t}{2} \quad \frac{{}^2_t(1 \quad )}{2} + \quad {}_t \quad \mathbb{E}[kg_t k_2^2] \\
&\quad + \frac{{}^2_t}{2} \quad \frac{d \ {}^2}{m} + \frac{d \ {}^2_0}{{}^2 m^2} \\
&= \quad \frac{}{2}\mathbb{E}[kg_t k_2^2] + \frac{{}^2}{2} \quad \frac{64L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{{}^2 m(m + \frac{{}^2_0}{\mathrm{b}} \ d)} \\
&\quad + \frac{64L^2 \quad \frac{{}^2_0}{\mathrm{b}} d \quad dT \log \frac{2:5T}{} \log \frac{2}{}}{{}^2 m^2 \quad \frac{m}{} + \frac{{}^2_0}{\mathrm{b}} d} + \frac{L^2 \ {}^2_0 d}{\mathrm{b} m^2}\Bigg)^1 \\
&\quad \text{—}(\mathbb{E}[L(w_{t\ 1})] \quad L \ ) \\
&\quad + \frac{{}^2}{2} \quad \frac{128L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2 \ {}^2} + \frac{L^2 \ {}^2_0 d}{\mathrm{b} m^2}\Bigg) \ :
\end{aligned}$$

Re-arranging the terms, we get

$$\begin{aligned}
\mathbb{E}[L(w_t)] \quad L \quad & 1 \quad \text{—} \ (\mathbb{E}[L(w_{t\ 1})] \quad L \ ) \\
& + \frac{{}^2}{2} \quad \frac{128L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2 \ {}^2} + \frac{L^2 \ {}^2_0 d}{\mathrm{b} m^2}\Bigg) \ :
\end{aligned}$$

Summing over $t = 1; \quad ; T$, we obtain

$$\begin{aligned}
\mathbb{E}[L(w_T)] \quad L \quad & 1 \quad \text{—} \ ^T (L(w_0) \quad L \ ) \\
& + \frac{T \ {}^2}{2} \quad \frac{128L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2 \ {}^2} + \frac{L^2 \ {}^2_0 d}{\mathrm{b} m^2}\Bigg) \ :
\end{aligned}$$

When $T > \frac{{}^2_0 \ {}^2}{128\mathrm{b} \log \frac{2}{}}$, it holds that

$$\frac{L^2 d \ {}^2_0}{\mathrm{b} m^2} < \frac{128L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2 \ {}^2};$$

and thus

$$\begin{aligned}
\mathbb{E}[L(w_T)] \quad L \quad & 1 \quad \text{—} \ ^T (L(w_0) \quad L \ ) \\
& + \frac{T \ {}^2}{} \quad \frac{128L^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2 \ {}^2}:
\end{aligned}$$

By taking

$$T = O\Bigg( \log \ \frac{m^2 \ {}^2}{L^2 d \log \frac{1}{}} \Bigg) = \log \ \frac{}{} \Bigg) \ ; \qquad (49)$$

we obtain that

$$\mathbb{E}[L(w_T)] \quad L \quad \mathcal{O}\Bigg( \frac{}{\log \frac{}{}} \Bigg)^2 \quad \frac{d \log^2 m \log \frac{1}{}}{m^2 \ {}^2}\Bigg)^1 :$$

For simplicity, we denote $c := \ - \ < 1$ and $F( \ ) := \frac{}{\log \frac{}{}} = \frac{}{\log \frac{1}{1 \ c}}$. Take the derivative of $F$ with respect to , we have

$$F^{\prime}( \ ) = \frac{\log \ 1 + \frac{c}{1 \ c} \quad \frac{c}{1 \ c}}{\log^2 \frac{1}{1 \ c}}:$$

By noting that $\frac{c}{1 \ c} > 0$ and using the inequality that $\log(1 + x) \quad x$ for $\forall x > \ 1$, we know that $F^{\prime}( \ ) \quad 0$. Thus $F( \ ) = \frac{}{\log \frac{}{}}$ monotonically decreases with . Moreover, since $0 \quad \frac{p}{d} \quad 1$, we know that for any possible $2 \ (0; 1]$, we have

$$\frac{1}{c} \quad 1 \quad \frac{1}{\log \frac{1}{1 \ c}} = F(1) \quad F( \ ) \quad \lim_{! \ 0} F( \ ) = \frac{1}{c}:$$

Therefore, we finally obtain that

$$\mathbb{E}[L(w_T)] \quad L \quad \mathcal{O}\Bigg( \frac{d \log^2 m \log \frac{1}{}}{m^2 \ {}^2} \Bigg) \ :$$

□

Theorem 4 reveals the trade-off between privacy, utility, and communication for OTA-FL under functions with PL condition. First of all, we show that our algorithm achieves the optimal dependence on the privacy budget/loss , device number $m$, and model dimension $d$. To this end, we establish the excess empirical risk lower bound in the following Theorem 5 for strongly-convex risk functions, which can be seen as a special case in the class of functions satisfying Polyak-Łojasiewicz condition.

**Theorem 5** (Excess empirical risk lower bound under PL condition)**.** Let $m; d \ 2 \ \mathbb{N}, \ > 0$, and $= o(1/m)$. For every device-level $( \ ; \ )$-DP FL algorithm, there is a FL setting with strongly convex $L( \ )$ such that, with probability at least $1/3$ (over the algorithm random coins), we must have

$$\mathbb{E}[L(w_T)] \quad L \quad \min \ 1; \frac{d}{m^2 \ {}^2} \qquad (50)$$

*Proof of Theorem 5.* Let $\grave{}(w; \ )$ be half the squared $\grave{}_2$-distance between $w$ and , that is

$$\grave{}(w; \ ) = \frac{1}{2} kw \quad k_2^2: \qquad (51)$$

For each local dataset $D_i = \{ \ _{i;1}; \quad ; \ _{i;n}\}$, the local empirical risk function is

$$L_i(w) = \frac{1}{2n} \sum_{j=1}^{n} \grave{}(w; \ _{i;j}): \qquad (52)$$

Define $\bar{} \ _i := \frac{1}{n} \sum_{j=1}^{n} \ _{i;j}$, then we can also write $L_i(w)$ as

$$L_i(w) = \frac{1}{2} kw \quad \bar{} \ _i k_2^2 + \frac{1}{2n} \sum_{j=1}^{n} k \ _i \quad \ _{i;j} k_2^2: \qquad (53)$$

The global empirical risk function $L( \ )$ is

$$L(w) = \frac{1}{2m} \sum_{i=1}^{m} kw \quad \bar{} \ _i k_2^2 + \frac{1}{2mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k \ _i \quad \ _{i;j} k_2^2: \qquad (54)$$

Define $\bar{} \ := \frac{1}{m} \sum_{i=1}^{m} \bar{} \ _i$. The minimizer of $L( \ )$ is . The excess empirical risk of parameter $w$ is

$$\begin{aligned}
L(w) \quad L \ &= \frac{1}{2m} \sum_{i=1}^{m} kw \quad \bar{} \ _i k_2^2 \quad \frac{1}{2m} \sum_{i=1}^{m} k \quad \bar{} \ _i k_2^2 \qquad (55) \\
&= \frac{1}{2} kw \quad k_2^2: \qquad (56)
\end{aligned}$$

Let $f_1; \ldots; _m g \quad f \; \frac{p}{d}; \frac{p}{d} g^d$. Then by directly using Part 2 of [41, Lemma 5.1], we can conclude that, for every device-level $(\ ;\ )$-DP algorithm, with probability at least $\frac{1}{3}$,

$$L(w) \quad L \qquad \min\ 1; \frac{d}{m^2\ ^2} \quad : \tag{57}$$

□

**Remark 6.** The impact of the limited communication band at edge to OTA-FL is dominated by the factors related to the compression ratio . We find that the utility bound is insensitive to , *i.e.*, the change of does not change the order of the learning utility bound. But the limit band indeed postpones the time it takes for the learning algorithm to achieve the best utility as indicated in the requirement of $T = O\ \log\ \frac{m^2\ ^2}{L^2 d \log \frac{1}{}} = \log$ —— where a smaller will give rise to more iterations for achieving the desired utility.

#### CASE 2: General smooth non-convex functions

We now consider general smooth non-convex global empirical risk function $L(\ )$. The utility of the algorithm is measured by the expected squared $\grave{}_2$ gradient-norm of the output model parameter, *i.e.*, $\mathsf{E}[k\, r\, L(w_{\text{output}})k_2^2]$.

**Theorem 6** (Utility upper bound for general smooth non–convex functions)**.** Suppose $L$ is -smooth and non-convex in general. In Algorithm 2, let be as (41), $_t = -$, $T > \frac{\frac{2}{0}\ ^2}{128\mathrm{b}\ \log\ ^2}$ and $T = O\ \ p\frac{m}{d \log \frac{1}{}}$ , then we have the following utility bound for $w_{\text{output}} = w_t$, where $t$ is uniformly sampled from $[T]$.

$$\mathsf{E}[k\, r\, L(w_{\text{output}})k_2^2] \quad \Theta\left(\frac{0\ \mathrm{q}\ \frac{}{d \log \frac{1}{}}\ 1}{m}\ \mathrm{A}\right); \tag{58}$$

where the $\Theta(\ )$ notation omits other poly-logarithmic terms and constants.

*Proof of Theorem 6.* Similar to the proof of Theorem 4, we have

$$\mathsf{E}[L(w_t) \quad L(w_{t\ 1})]$$
$$\frac{}{2}\mathsf{E}[k\, r\, L(w_{t\ 1})k_2^2] + \frac{128L^2\ ^2 dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2\ ^2}:$$

From this, we get

$$\mathsf{E}[k\, r\, L(w_{t\ 1})k_2^2] \tag{59}$$
$$\frac{2\ \mathsf{E}[L(w_{t\ 1}) \quad L(w_t)]}{} + \frac{256L^2\ \ dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2\ ^2} \tag{60}$$

Summing over $t = 1; \quad ; T$ and taking the average, we obtain

$$\frac{1}{T}\sum_{t=1}^{\bar{X}} \mathsf{E}[k\, r\, L(w_{t\ 1})k_2^2] \tag{61}$$
$$\frac{2\ (L(w_0) \quad L\ )}{T} + \frac{256L^2\ \ dT \log \frac{2:5T}{} \log \frac{2}{}}{m^2\ ^2}: \tag{62}$$

By $T = O\ \ p\frac{m}{d \log \frac{1}{}}$ , we have

$$\mathsf{E}[k\, r\, L(w_{\text{output}})k_2^2] \tag{63}$$
$$= \frac{1}{T}\sum_{t=1}^{\bar{X}} \mathsf{E}[k\, r\, L(w_{t\ 1})k_2^2] \quad \Theta\left(\frac{0\ \mathrm{q}\ \frac{}{d \log \frac{1}{}}\ 1}{m}\ \mathrm{A}\right): \tag{64}$$

□

**Remark 7.** Our utility upper bound of $\Theta\ \frac{p\ \frac{}{d \log \frac{1}{}}}{m}$ for general non-convex loss coincides with the state-of-the-art device-level DP utility bound presented in [42] for conventional federated learning without AirComp. Moreover, we can observe that the limited communication band, captured by the compression ratio , has a similar effect as we derived for the case of objective function with PL condition. That is, the limited band indeed delays the convergence time of the learning algorithm to attain the optimal utility as specified by the condition of $T = O\ \ p\frac{m}{d \log \frac{1}{}}$ , where a smaller implies more iterations to achieve the desired utility.

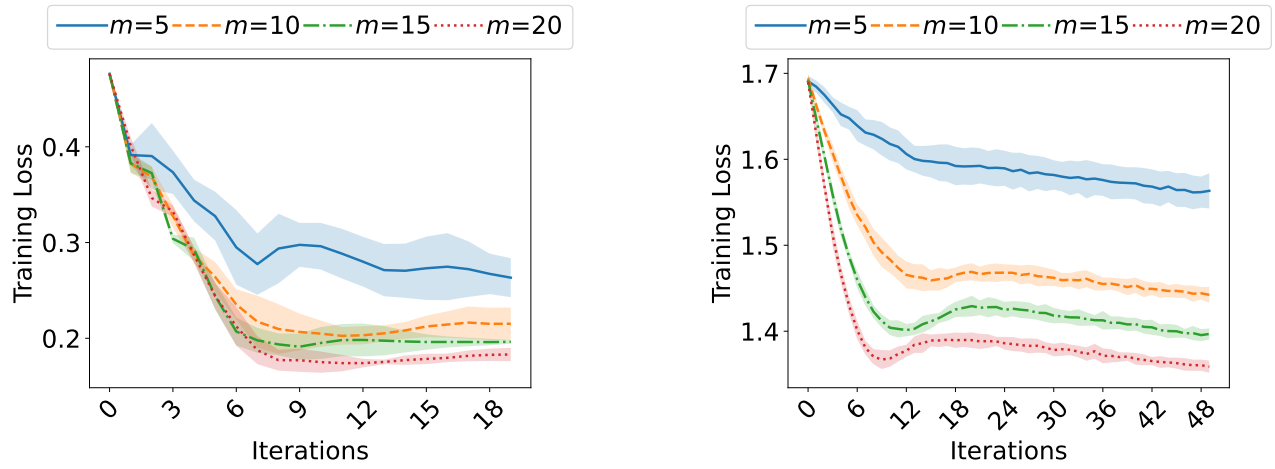
## 5 Experiments

In this section, we evaluate the performance of PROBE using real-world benchmark datasets. We will first showcase the convergence accuracy and efficiency of PROBE in different settings with varying system scales, privacy budgets, and compression ratios, respectively. Then, we will also illustrate the communication efficiency and resilience against CSI attacks for PROBE. Finally, We compare PROBE with the state-of-the-art baseline method called DPRP-FedSGD proposed in [35].

### 5.1 Experiment Setup

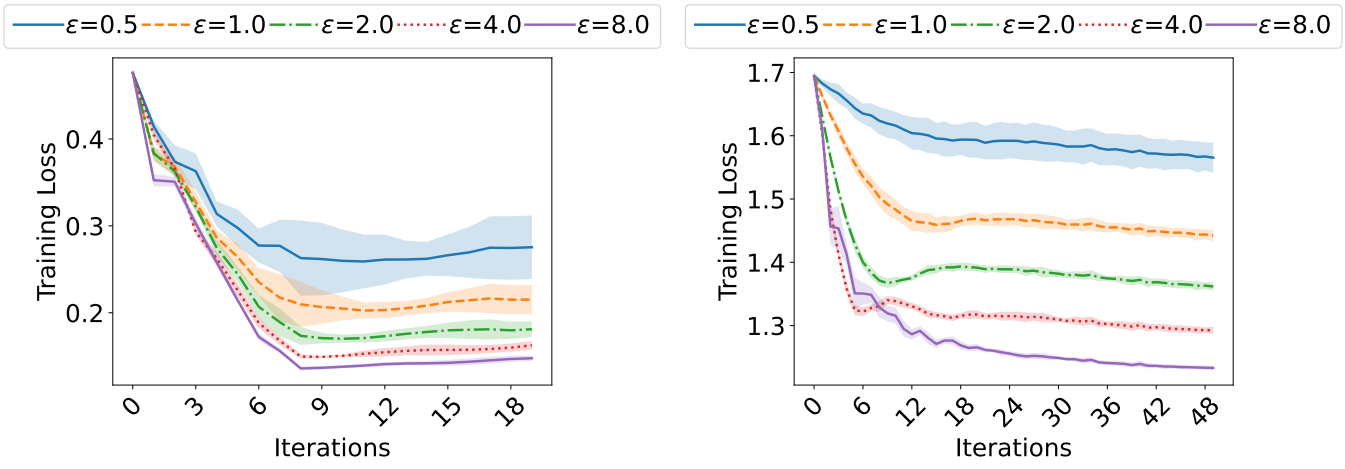
We apply PROBE to non-convex learning tasks on two real-world benchmark datasets, MNIST [43] and CIFAR-10 [44], using the 18-layer residual network (ResNet-18) [45] model for image classification. We estimate the problem-related parameters such as based on the dataset and the model following the methods in [46], [47]. We fix $= 10^{\ 3}$, $_0 = 1{:}0$, and $c_i = 0{:}8$ for $8i\ 2\ [m]$ throughout the experiments. We set the number of iterations $T$ to 20 and 50 for MNIST and CIFAR-10, respectively. We assume that the devices have power constraints $P_i$'s uniformly distributed in $[25; 30]$. We use $= 1{:}0$, $= 0{:}8$, $= 0{:}8$ as the default setting for PROBE and vary one parameter at a time to evaluate the effects of different factors. We measure the algorithmic convergence by the training loss. We repeat each experiment at least 10 times and report the average results in the figures.

For comparison with the state-of-the-art methods, we use the DPRP-FedSGD algorithm from [35] as the baseline method, which to the best of our knowledge is the only existing work that considered differentially private OTA-FL with communication compression. Specifically, both DPRP-FedSGD and our algorithm PROBE are based on applying federated stochastic gradient descent for minimizing the given loss function based on the local gradients. The main difference is that DPRP-FedSGD uses Johnson-Lindenstrauss (JL) random projection for reducing the dimension of the local updates and uses a different AirComp signal transmission strategy as we discussed in Remark 2.




(a) MNIST (b) CIFAR-10

Fig. 1: Training loss versus iterations for different system scales on MNIST and CIFAR-10 datasets. The results show that PROBE achieves better convergence accuracy and efficiency with the increased number of devices.




(a) MNIST (b) CIFAR-10

Fig. 2: The effect of privacy budget on the performance of PROBE on MNIST and CIFAR-10 datasets. Larger privacy budgets lead to lower training losses, but the improvement becomes marginal after = 2:0.

Specifically, for each local gradient $g_{i;t}$, DPRP-FedSGD reduces its dimension using the following gradient projection step:

$$\mathbf{g}_{i;t} = p\frac{1}{d}\ \mathsf{D}\ \ \mathsf{U}\ g_{i;t}; \tag{65}$$

where $\mathsf{D}$ is a ( $d$) $d$ rectangular diagonal matrix, *i.e.*, $[\mathsf{D}]_{i;i} = 1; i\ 2\ [\ \ d]$ and $[\mathsf{D}]_{i;j} = 0; i\ \mathbf{6}\ j$, and $\mathsf{U}\ 2\ \mathsf{R}^{d\ \ d}$ is a random projection matrix generated with entries drawn independently from Rademacher distribution (symmetric Bernolli taking values +1 and 1 with probability $\frac{1}{2}$), or from the Gaussian distribution of zero mean and unit variance as $[\mathsf{U}]_{i;j}$ $N(0;1)$, or from Achlioptas distribution, given by

$$[\mathsf{U}]_{i;j} = \begin{cases} +p\overline{s}; & \text{with probability } \frac{1}{2s} \\ 0; & \text{with probability } 1\ \ \frac{1}{s} \\ p\overline{s}; & \text{with probability } \frac{1}{2s}: \end{cases} \tag{66}$$

After that, DPRP-FedSGD sets the vector to be sent by each device as

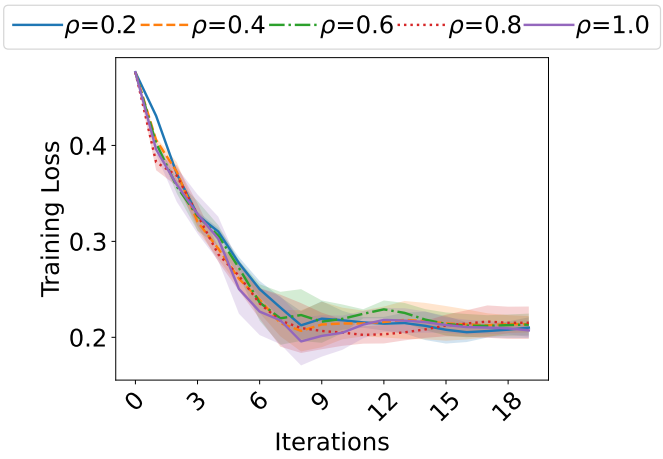
$$x_{i;t} = \frac{p\ \overline{\ _{i;1}P_i}}{L}\mathbf{g}_{i;t} + q\ \overline{\ _{i;2}P_i}\ \ _{i;t}; \tag{67}$$

where $_{i;1}\ 2\ [0;1]$ and $_{i;2}\ 2\ [0;1\ \ \ _{i;1}]$ denote the fraction of power allocated to the normalized local gradient vector $\frac{1}{L}\ g_{i;t}$ and the DP noise $_{i;t}$ $N(0;1\ \ \mathsf{I}_d)$, respectively. The power allocation parameters satisfy $_{i;1} + \ _{i;2}$ 1 so that the maximum power constraint of $P_i$ is satisfied. In order to form an unbiased estimate of the global gradient, all devices pick the $_{i;1}$'s as:

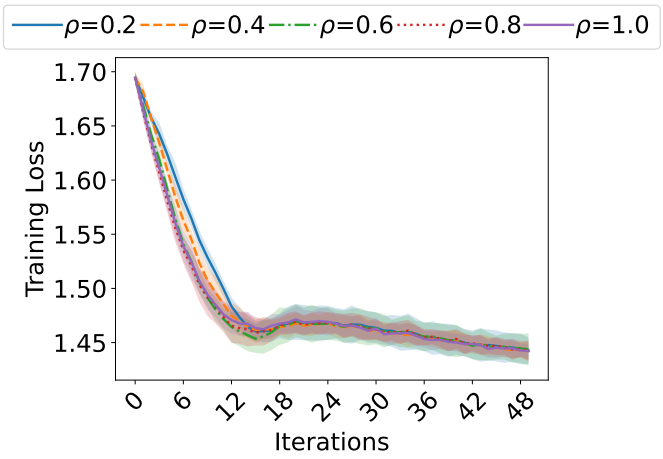
$$_{i;1} = \frac{\min_{i^0 2[m]}\ \ i^0}{i} = \frac{0}{i}; \tag{68}$$

so that the server can receive an aggregated vector where all local gradients have the same weight:

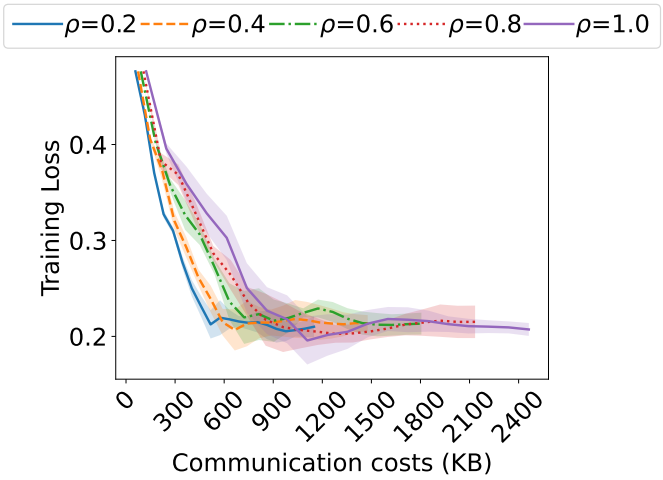
$$y_t = \frac{p\ \overline{\min_{i^0 2[m]}\ \ i^0}}{L}\sum_{i=1}^{N} g_{i;t} + \sum_{i=1}^{N} q\ \overline{\ _{i;2}P_i}\ \ _{i;t} + z_t: \tag{69}$$



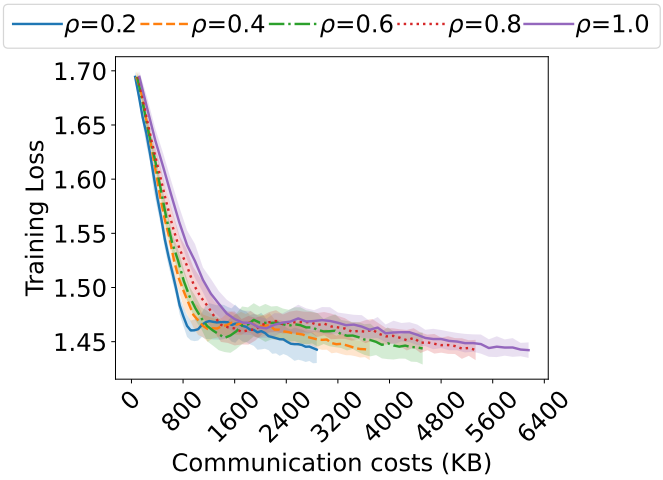

(a) MNIST: Loss v.s. Iterations




(b) CIFAR-10: Loss v.s. Iterations




(c) MNIST: Loss v.s. Communication Costs




(d) CIFAR-10: Loss v.s. Communication Costs

Fig. 3: The effect of compression ratio on the performance of PROBE on MNIST and CIFAR-10 datasets. Smaller compression ratios lead to slightly slower convergence but better communication efficiency with almost the same training loss.

After that, DPRP-FedSGD updates the global model using the gradient descent step just like our PROBE.

We will conduct two experiments for the comparison with DPRP-FedSGD. Firstly, as we mentioned in Remark 2, for each device, its per-round privacy guarantee offered by DPRP-FedSGD is subject to the ratio of its individual effective SNR and the worst effective SNR across devices and the privacy loss increases when the effective SNR becomes more consistent. For the experiment, we will investigate how the per-round privacy loss of the device with the largest effective SNR $_{max}$ varies when the ratio of $_{max}/_{min}$ changes for both algorithms. Specifically, we fix $_{min} = 12$ and change the value of $_{max}$ such that $_{max}/_{min}$ varies in this experiment. Next, we study the privacy and communication trade-offs achieved by the two algorithms. We will let the compression ratio vary and present the privacy loss under different compression levels.

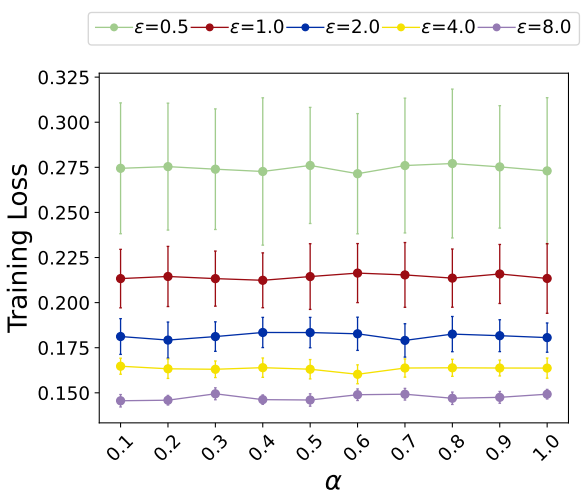
## 5.2 Results and Discussions

### 5.2.1 Impact of system scale

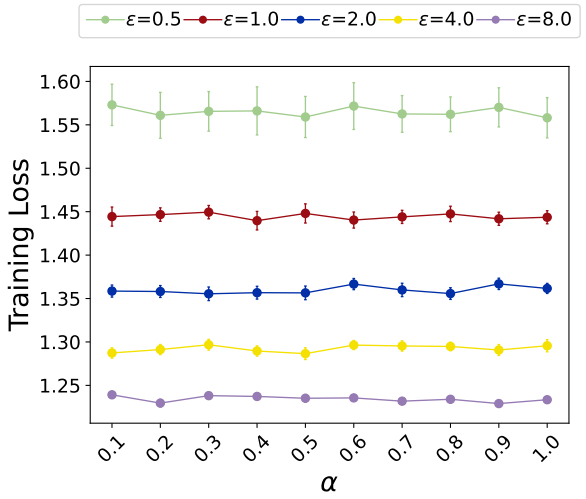
We evaluate the effect of the device number $m$ on the performance of PROBE by varying $m$ from 5 to 20 with step 5. The results are shown in Fig. 1. It can be observed that, as the number of devices increases, the training loss decreases more rapidly and stabilizes at a lower value for both MNIST and CIFAR-10 datasets. This trend is consistent with our theoretical results (Theorem 4 and Theorem 6) and the experimental findings in [26]. For instance, with CIFAR-10, the training loss for $m = 20$ is significantly lower than that for $m = 5$, indicating enhanced performance with increased system scale. This is due to the privacy amplification effect of AirComp. As $m$ increases, each device injects less Gaussian noise as stated in Theorem 3, which results in a more accurate global gradient estimate as proved in Lemma 2. Hence, PROBE can leverage the increased system scale to achieve better convergence accuracy and efficiency.

### 5.2.2 Impact of privacy budget

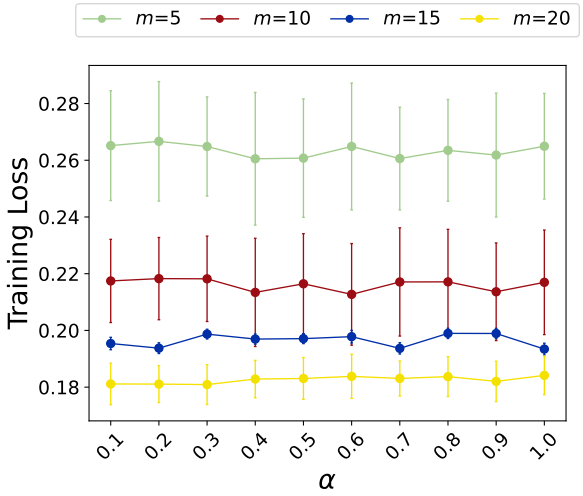
We examine the effect of the privacy budget on the performance of PROBE by increasing from 0.5 to 8.0 with a factor of 2. Fig. 2 shows the results. We notice that, for both MNIST and CIFAR-10 datasets, the final training loss value after stabilization reduces as the privacy budget increases. This is because a higher implies less noise addition for each



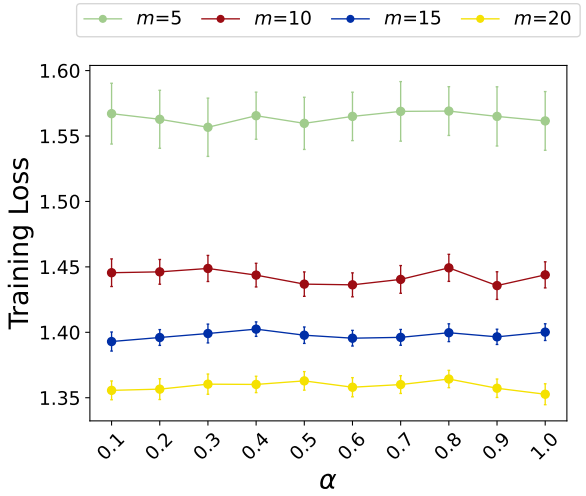

(a) MNIST: Results under different privacy budget $\epsilon$




(b) CIFAR-10: Results under different privacy budget $\epsilon$

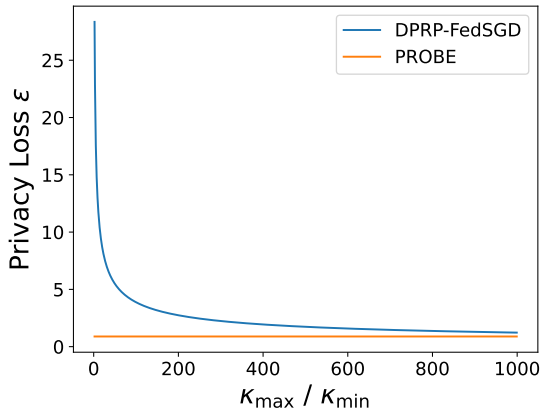



(c) MNIST: Results under different system scale $m$

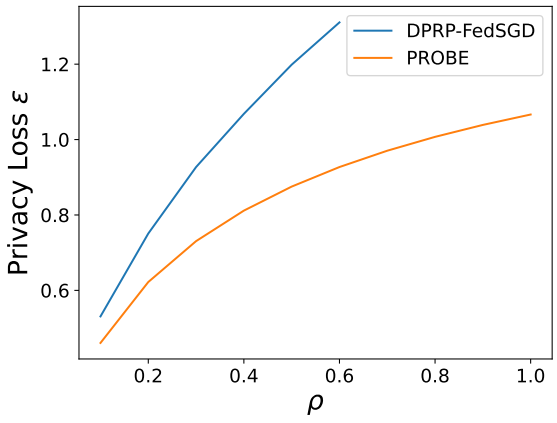



(d) CIFAR-10: Results under different system scale $m$

Fig. 4: The robustness of PROBE to CSI attacks on MNIST and CIFAR-10 datasets. The final training loss is almost unaffected by different values, indicating the immunity of PROBE to CSI attacks. The final training loss gets lower with a larger privacy budget or a larger system scale.




(a) Privacy loss incurred by different $\kappa_{\max}/\kappa_{\min}$




(b) Privacy loss under different compression levels

Fig. 5: Results for comparison of PROBE with the baseline method DPRP-FedSGD.

device, which improves the accuracy of the global gradient estimate as proved in Lemma 2. Moreover, with the increase of privacy budget , the magnitude of the final training loss decrease also gets smaller, which implies the final training loss gets less affected by the change of when is large. In particular, the training loss decrease when the privacy budget exceeds 2 is considerably smaller than that when is below 2, indicating that PROBE can balance privacy and accuracy with a moderate privacy budget. Finally, the results also indicate that PROBE is more affected by the privacy budget on CIFAR-10 than on MNIST, which may be related to the higher complexity and diversity of the CIFAR-10 dataset. Therefore, the optimal privacy budget may vary depending on the dataset characteristics and the required level of privacy protection.

#### 5.2.3 Impact of compression ratio

To study the influence of the compression on the performance, we select different ranging from 0.2 to 1.0 with step 0.2 and run PROBE respectively. The results are shown in Fig. 3. We can see that, from the view of training loss versus iterations, smaller values lead to slightly slower convergence but the training losses under different compression ratios are almost the same for both MNIST and CIFAR-10 datasets. This suggests that PROBE is robust to the compression ratio and can achieve similar accuracy with different values. However, from the view of training loss versus communication costs, smaller values cost less communication costs to achieve the same training loss. This indicates that a lower compression ratio can improve the communication efficiency of PROBE to make it adapted to limited band in practice, but at the expense of slower convergence. Such an observation corroborates with our previous discussion given in Remark 6 and Remark 7. Therefore, there is a trade-off between communication overheads and the convergence rate of PROBE depending on the band limit.

#### 5.2.4 Resilience against CSI attack

We test the resilience of PROBE against CSI attacks by varying the scaling parameter from 0.1 (the most severe attack) to 1.0 (no attack) with step 0.1. We show the final training loss versus under different privacy budget or system scale $m$ on both datasets in Fig. 4. We observe that, for both MNIST and CIFAR-10 datasets, the final training loss is almost invariant to different values under different settings of and $m$. This implies that PROBE is robust to CSI attacks and can preserve its performance regardless of the value of . Moreover, we also notice that the final training loss decreases when the privacy budget is larger or the system scale $m$ is larger, which is consistent with our previous results and conclusions. Therefore, PROBE can withstand different levels of CSI attacks across different settings with different privacy budgets or system scales.

#### 5.2.5 Comparison with Baseline

We compare the performance of our proposed algorithm PROBE with the state-of-the-art baseline method DPRP-FedSGD on the privacy and communication trade-offs in OTA-FL. The results are shown in Figure 5. In Figure 5(a), we plot the privacy loss incurred by different $_{max}=_{min}$ ratios. The y-axis shows the per-round privacy loss and the x-axis shows the $_{max}=_{min}$ ratio. We can see that PROBE achieves significantly lower privacy loss than DPRP-FedSGD across all values of $_{max}=_{min}$. This demonstrates the robustness of PROBE against the heterogeneity of the effective SNRs among devices. In contrast, DPRP-FedSGD suffers from high privacy loss when the effective SNRs are more consistent, which is consistent with what we discussed in Remark 2. In Figure 5(b), we plot the privacy loss under different compression levels. The y-axis shows the privacy loss and the x-axis shows the compression ratio . We can see that both algorithms incur higher privacy loss as increases, which is expected due to the information loss caused by compression. However, PROBE still outperforms DPRP-FedSGD at all compression levels, indicating that PROBE can achieve better privacy and communication trade-offs than DPRP-FedSGD.

## 6 Conclusions

In this paper, we proposed a private and communication-efficient framework for over-the-air federated learning at edge. To achieve this, we designed a locally differentially private mechanism that integrates Gaussian masking perturbation and random sparsification techniques. By utilizing common randomness, the privacy amplification offered by sparsification and AirComp successfully adds up, which provides the best possible local privacy guarantees for each edge device. Furthermore, our framework requires lightweight computation, respects specific power constraints of all devices, and is immune to potential CSI attacks. All these features make the proposed framework implement well in reality. For future work, it is interesting to explore the interplay of privacy preservation with other communication-efficient techniques for OTA-FL.

## References


[1] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 8 2019.

[2] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.

[3] F. Dressler, C. F. Chiasserini, F. H. P. Fitzek, H. Karl, R. Lo Cigno, A. Capone, C. E. Casetti, F. Malandrino, V. Mancuso, F. Klingler, and G. A. Rizzo, "V-Edge: Virtual Edge Computing as an Enabler for Novel Microservices and Cooperative Computing," *IEEE Network*, vol. 36, no. 3, pp. 24–31, 5 2022.

[4] N. Chen, T. Qiu, L. Zhao, X. Zhou, and H. Ning, "Edge intelligent networking optimization for internet of things in smart city," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 26–31, 2021.

[5] S. Pandya, G. Srivastava, R. Jhaveri, M. R. Babu, S. Bhattacharya, P. K. R. Maddikunta, S. Mastorakis, M. J. Piran, and T. R. Gadekallu, "Federated learning for smart cities: A comprehensive survey," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102987, 2023.

[6] Y. Guo, F. Liu, Z. Cai, L. Chen, and N. Xiao, "Feel: A federated edge learning system for efficient and privacy-preserving mobile healthcare," in *Proceedings of the 49th International Conference on Parallel Processing*, 2020, pp. 1–11.

[7] Z. Lian, Q. Yang, W. Wang, Q. Zeng, M. Alazab, H. Zhao, and C. Su, "Deep-fel: Decentralized, efficient and privacy-enhanced federated edge learning for healthcare cyber physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 3558–3569, 2022.

[8] B. Yang, X. Cao, K. Xiong, C. Yuen, Y. L. Guan, S. Leng, L. Qian, and Z. Han, "Edge intelligence for autonomous driving in 6g wireless system: Design challenges and solutions," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 40–47, 2021.

[9] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8423–8434, 2021.

[10] 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," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data *et al.*, "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[12] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Transactions on information theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[13] B. Xiao, X. Yu, W. Ni, X. Wang, and H. V. Poor, "Over-the-air federated learning: Status quo, open challenges, and future directions," *arXiv preprint arXiv:2307.00974*, 2023.

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

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

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

[17] ——, "A sequential gradient-based multiple access for distributed learning over fading channels," in *Proceedings of the conference on Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 303–307.

[18] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *Proceedings of the International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[19] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proceedings of the Symposium on Security and Privacy (S&P)*. IEEE, 2019, pp. 691–706.

[20] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proceedings of the Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 3–18.

[21] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.

[22] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *Proceedings of the International Conference on Computer Communications (INFOCOM)*. IEEE, 2019, pp. 2512–2520.

[23] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14 774–14 784, 2019.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the conference on Theory of Cryptography Conference (TCC)*. Springer, 2006, pp. 265–284.

[25] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2020.

[26] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proceedings of the International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2604–2609.

[27] B. Hasırcıoğlu and D. Gündüz, "Private wireless federated learning with anonymous over-the-air computation," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5195–5199.

[28] M. Wang, W. Fu, X. He, S. Hao, and X. Wu, "A survey on large-scale machine learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2574–2594, 2020.

[29] X. Huang, P. Li, H. Du, J. Kang, D. Niyato, D. I. Kim, and Y. Wu, "Federated learning-empowered ai-generated content in wireless networks," *IEEE Network*, 2024.

[30] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, D. I. Kim, V. Leung *et al.*, "Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services," *arXiv preprint arXiv:2303.16129*, 2023.

[31] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A survey on chatgpt: Ai-generated contents, challenges, and solutions," *arXiv preprint arXiv:2305.18339*, 2023.

[32] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

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

[34] J. Zhang, N. Li, and M. Dedeoglu, "Federated learning over wireless networks: A band-limited coordinated descent approach," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[35] A. Sonee, S. Rini, and Y.-C. Huang, "Wireless federated learning with limited communication and differential privacy," in *Proceedings of the conference on Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 01–06.

[36] R. Hu, Y. Gong, and Y. Guo, "Federated learning with sparsification-amplified privacy and adaptive optimization," *arXiv preprint arXiv:2008.01558*, 2020.

[37] X. Zhang, M. Fang, J. Liu, and Z. Zhu, "Private and communication-efficient edge learning: a sparse differential gaussian-masking distributed sgd approach," in *Proceedings of the International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, 2020, pp. 261–270.

[38] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017, pp. 1273–1282.

[39] 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.

[40] R. Hu, Y. Guo, and Y. Gong, "Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy," *IEEE Transactions on Mobile Computing*, 2023.

[41] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2014, pp. 464–473.

[42] A. Cheng, P. Wang, X. S. Zhang, and J. Cheng, "Differentially private federated learning with local regularization and sparsification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 122–10 131.

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

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

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] 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 journal on selected areas in communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[47] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas, "Cost-effective federated learning design," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.