# Random Aggregate Beamforming for Over-the-Air Federated Learning in Large-Scale Networks

Chunmei Xu, *Student Member, IEEE*, Shengheng Liu, *Member, IEEE*, Yongming Huang, *Senior Member, IEEE*, Björn Ottersten, *Fellow, IEEE*, and Dusit Niyato, *Fellow, IEEE*

## Abstract

At present, there is a trend to deploy ubiquitous artificial intelligence (AI) applications at the edge of the network. As a promising framework that enables secure edge intelligence, federated learning (FL) has received widespread attention, and over-the-air computing (AirComp) has been integrated to further improve the communication efficiency. In this paper, we consider a joint device selection and aggregate beamforming design with the objectives of minimizing the aggregate error and maximizing the number of selected devices. This yields a combinatorial problem, which is difficult to solve especially in large-scale networks. To tackle the problems in a cost-effective manner, we propose a random aggregate beamforming-based scheme, which generates the aggregator beamforming vector via random sampling rather than optimization. The implementation of the proposed scheme does not require the channel estimation. We additionally use asymptotic analysis to study the obtained aggregate error and the number

Manuscript received XXX XX, 2021; revised XXX XX, XXXX; accepted XXX XX, XXXX. Date of publication XXX XX, XXXX; date of current version XXX XX, XXXX. This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62001103 and 61720106003, the National Key R&D Program of China under Grant Nos. 2018YFB1800801 and 2020YFB1806608. (*Corresponding authors: S. Liu and Y. Huang*.)

C. Xu, S. Liu and Y. Huang are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: {xuchunmei; s.liu; huangym}@seu.edu.cn).

B. Ottersten is with Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, L-1359 Luxembourg (e-mail: bjorn.ottersten@uni.lu).

D. Niyato is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: dniyato@ntu.edu.sg).

of the selected devices when the number of devices becomes large. Furthermore, a refined method that runs with multiple randomizations is also proposed for performance improvement. Extensive simulation results are presented to demonstrate the effectiveness of the proposed random aggregate beamforming-based scheme as well as the refined method.

**Index Terms**

Federated learning, over-the-air computation (AirComp), device selection, aggregate beamforming, large-scale distributed systems.

## I. Introduction

The ubiquitous mobile gadgets and Internet of Things (IoT) devices, together with the recent revival and breakthrough of artificial intelligence (AI), have inspired people to envision intelligence at the edge of the wireless networks [1], [2]. Many AI tasks are computationally intensive, which are conventionally trained at a powerful server center with a large amount of data collected and stored. However, the centralized-training paradigm is generally connected to high latency, because of the big volume of data to be transmitted. In addition, collecting sensitive data such as geographic locations [3] and personal health records [4] can lead to serious privacy violations. Fortunately, mobile edge computing (MEC) brings storage and computation resources closer to the user and allows circulation of data to take place locally [5]. Moreover, the latest mobile devices have already been equipped with high-performance graphical-processing units (GPUs), which can be used for task training. Riding on these trends, the deployment of AI algorithms at the network edge is feasible. On the other hand, to address the privacy concerns in distributed learning, the framework of federated learning (FL) is surging in popularity, which collaboratively trains a globally shared model without sharing the sensitive raw data [6]. Despite effective prevention of data leakage, communication overhead is a principal bottleneck of the FL schemes, since frequent model transmission via wireless links are required [1], [7], [8].

It is widely foreseen that the future wireless network service such as IoT will leverage dense edge devices providing pervasive sensing and actuation ability. For a large-scale system with a massive number of devices, normally only a small fraction of devices are selected for local model updating. The reasons are twofold respectively from machine learning and communication perspectives. First, it is observed that the learning performance may degrade when excessive devices participate in the global model aggregation [6], as too large sized batches lead to higher validation error [9]. Second, aggregation is constrained by the latency requirement and available wireless resources, which restricts the number of devices involved. Hence, elaborate design of device selection and transmission mechanism is of paramount importance for edge intelligence in future large-scale wireless networks.

For device selection, the original protocol was designed to randomly select a fraction of devices for local model updating of FL [6]. However, this approach is inefficient in practical settings, since heterogeneous devices have different reliability values, data sizes, computation/storage capacities, and experience distinct wireless channels. For instance, unreliable devices may perform malicious updates, which degrades the learning performance, as FL is susceptible to adversarial attacks [10], [11]. Besides, the *stragglers* with low storage and computation capacities or appalling channel conditions generally require a longer time to upload their local models. In a synchronized FL system, the aggregator has to wait or ignore the stragglers, which in return impairs the learning efficiency [12], [13]. In this context, three simple scheduling policies were implemented in [14] and the corresponding FL convergence rates were analyzed by taking into account the wireless channel conditions [14]. Further, Nishio et al. [15] consider maximizing the number of the selected devices under the training time budget constraints, which requires the collection of the available resource information of the devices. However, device selection schemes in the above studies are not assisted by proper transmission mechanism that can further improves the communication efficiency.

Improving the communication efficiency between the selected devices and the aggregator helps reduce the communication and training time budgets. Previous related works have considered reducing the time budgets by cutting down the size of the transmitted data via data quantization [16] or model sparsification [17]. Stich et al. also proposed to aggregate the local models less frequently, where the global model is attained after the local models are updated for multiple steps rather than after each iteration [18]. Additionally, the communication efficiency can be improved by lifting the transmission rate via advanced communication technologies [19]. Conventionally, in data aggregation, the receivers need to decode each individual transmitted data before further computations. This pattern is commonly referred to as the separated-communication-and-computation principle. Based on this conventional principle, the problem of joint device selection and bandwidth allocation is addressed [20], where the convergence rate is maximized under the time budget constraint determined by the achievable computation latency. As mitigating the generated interferences consumes more wireless resources, this scheme is inefficient especially given the limited resources. As an alternative, the emerging over-the-air computation (AirComp) principle [21], [22] integrates computation and communication by leveraging the waveform superposition property [19], [23]–[25]. The over-the-air analog aggregation can significantly improve the communication efficiency compared to the digital manner [19], [23]

Nevertheless, AirComp introduces signal distortion due to the fading channels and noises, which can be measured by mean square error (MSE). Such error deviates the aggregate data from the desired one, which harms FL learning performance as well as the communication efficiency [19]. Specifically, the existence of error implies that the received model parameters at the aggregator are perturbed. A large perturbation can lead to divergence of the training loss and degrade the classification/regression performance [26]. Although error control in the upper layer can be applied, it inevitably lowers the communication efficiency. Thus, the aggregate error must be mitigated. To this end, power allocation for over-the-air FL is considered [27] by taking gradient statistics into account. Utilizing the multiple antennas equipped at the aggregator, the aggregate beamforming vector and the scaling factor were designed by incorporating dynamic learning rates [25]. Zhu et al. [19] considered joint device selection and transmit power design and proposed the method based on the derived communication-and-learning tradeoffs. Essentially, the optimization of transmit power, aggregate beamforming and other resource allocation targeting at MSE reduction is to align the signals from the selected devices. On the other hand, a moderate perturbation on the model parameters is beneficial. It avoids over-fitting and improves the robustness of a neural network against adversarial attacks, which can be regarded as a regulation technique [28]–[31]. In this regard, selecting suitably more devices for local model updates and global model aggregation can improve the learning efficiency, since large-sized equivalent minibatch speeds up the training process while guaranteeing the achieved learning performance [9], [20]. Aiming at maximizing the number of the selected devices, Yang et al. [24] considered the problem of joint device selection and aggregate beamforming vector optimization under the aggregate error constraint by using the difference of convex (DC).

Design of joint device selection and transmission scheme for massive number of devices deploying AI applications is a meaningful topic. As the underlying combinatorial problems in the large-scale systems are extremely difficult to solve, it is challenging and, to the best of our knowledge, remains unexplored yet. The size of the selected device subset grows exponentially with the number of devices, which makes exhaustive search prohibitive. Despite the simplicity of the random device selection protocol, the resulting aggregate error can be unacceptable. As a compromise, conventional optimization methods can achieve good performance, but the required computational complexity may still be high. Besides, the acquisition of channel information in a large-scale system induces extremely high estimation overhead. Therefore, a new cost-effective method that can still achieve adequate performance is required. In this paper, we investigate the joint device selection and aggregate beamforming design in a large-scale system (e.g., IoT), where the FL framework and the AirComp technique are adopted. The heterogeneity of the devices focuses on the wireless channel states. Two relevant objectives, i.e., MSE minimization with a fixed number of devices, and the number of selected devices maximization under the MSE constrain, are considered for the systems with different requirements/goals, which make this paper comprehensive. The contributions of this paper are elaborated as follows:

- We propose a random aggregate beamforming-based scheme to solve the two combinatorial problems in a large-scale system. The core idea is to uniformly sample a vector from the complex unit sphere as the aggregate beamforming vector first, and select the devices afterwards. The implementation complexity of the proposed methods is low, which does not require the channel state information.
- Through asymptotic analysis, we prove that the minimum aggregate error can be approached using the proposed methods when the number of devices becomes large. We also derive the number interval as well as the average value for the problem of maximizing the number of the selected devices.
- To improve the obtained performance of both the problems in practical scenarios where the number of devices is less than infinity, we propose the refined methods that samples multiple vectors from the complex unit sphere. Although the required number of the vectors that achieves acceptable performance cannot be explicitly given, it provides an insight to obtain solutions with performance improvement.

The remainder of the paper is organized as follows. Section II gives the system model, where FL and AirComp technique are detailed. Section III formulates the investigated problems including the minimization of the aggregate error and the maximization of the number of the selected devices. In Section IV, we propose the aggregate beamforming based methods for both considered problems and discuss their theoretical performance, and the refined methods are given in Section V for performance improvement. Simulation results are presented in Section VI and the paper is concluded in Section VII.

## II. System Model

We consider a large-scale wireless network with edge intelligence, which consists of $K$ single-antenna devices and an aggregator equipped with $N \ll K$ antennas. Each device $k$ owns a local dataset $\mathcal{D}_k$, which collectively constitutes the global dataset $\mathcal{D} = \cup_{k\in\mathcal{K}}\mathcal{D}_k$. Typically, the objective function of a learning task is to find the model $\mathbf{w}^o$ that minimizes the loss function $P\left(\mathbf{w};\mathcal{D}\right)$, where $\mathbf{w} \in \mathbb{R}^D$ is the model weight parameters with the dimension of $D$. Although the centralized gradient decent method can be applied to update the model $\mathbf{w}$ in order to obtain $\mathbf{w}^o$, it becomes impractical as local datasets $\mathcal{D}_k$ at each device cannot be accessed by others due to the privacy and latency concerns. To handle these concerns, the FL can be adopted.

The iterative learning process of FL includes: 1) A subset of devices are selected; 2) The selected devices update the model with local data, and upload the updated model back to the aggregator; 3) The aggregator receives these local models and aggregates them to obtain the global model. As shown in Fig. 1, each device $k$ updates its model based on the locally stored dataset $\mathcal{D}_k$, and transmits the local model back to the aggregator. Then, the global model is updated by averaging the local models from the participating devices, which is referred to as the model aggregation phase. Following the updating rules of FL, the local models $\mathbf{w}_k, k \in \mathcal{K}$ and the global model $\mathbf{w}$ are respectively updated by

$$\mathbf{w}_k^{i+1} = \mathbf{w}^i - \mu \mathbf{g}_k\left(\mathbf{w}^i\right), \tag{1}$$

and

$$\mathbf{w}^{i+1} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k^{i+1}, \tag{2}$$

where superposition $i$ is the iteration index, $\mu$ is the learning rate or step size. $\mathbf{g}_k(\mathbf{w}^i) = \nabla_{\mathbf{w}^i} P_k\left(\mathbf{w}^i\right)$ is the gradients of the loss on the local dataset $\mathcal{D}_k$ with respect to (w.r.t.) $\mathbf{w}^i$, where $P_k\left(\mathbf{w}^i\right) = \frac{1}{|\mathcal{D}_k|}\sum_{n=1}^{|\mathcal{D}_k|} Q\left(\mathbf{w}^i;\mathcal{D}_k^n\right)$ is the loss function with $Q\left(\mathbf{w}^i;\mathcal{D}_k^n\right)$ the loss value on the $n$-th data sample $\mathcal{D}_k^n$.

Apart from the observation from [6], the limited wireless resources also deter the participation of all devices in updating the global model. Thus, only a subset of devices are selected for local models updating and the global model aggregation. In the considered model, AirComp technique is applied for the model aggregation in order to improve the communication efficiency. The transmit symbol vector of device $k$ at the $i$-th iteration is customized as $\mathbf{s}_k^i \triangleq \mathbf{w}_k^i$, which is assumed to be normalized with unit variance, i.e., $\mathbb{E}\left[\mathbf{s}_k^i\left(\mathbf{s}_k^i\right)^{\mathrm{H}}\right] = \mathbf{I}$. For notational convenience,

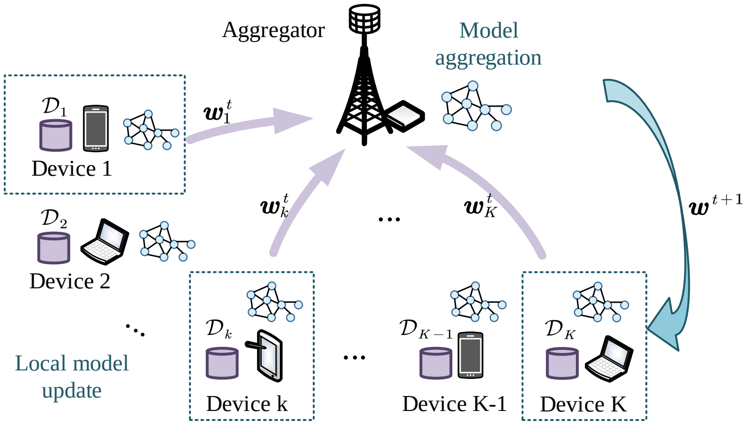


Figure 1: System hierarchy model under investigation.

the $d$-th element of $\mathbf{s}_k^i$, $\mathbf{w}^i$ and $\mathbf{g}_k$, i.e., $\mathbf{s}_k^i[d]$, $\mathbf{w}^i\left[d\right]$ and $\mathbf{g}_k[d]$, are denoted as $s_k$, $w$ and $g_k$. By denoting the set of selected device as $\mathcal{S}$, the desired signal based on (2) is written as

$$y_{\text{des}} = \sum_{k \in \mathcal{S}} w_k = \sum_{k \in \mathcal{S}} s_k. \tag{3}$$

In implementing AirComp, the symbols $s_k$ are modulated in an analog manner and precoded by the transmit coefficients $b_k$. The amplitude of $b_k$ means the transmit power of device $k$ and its phase is used to help align the obtained signals at the aggregator. Then, these signals are superimposed over the air, and the received signals at the aggregator are combined by the aggregate beamforming vector $\mathbf{m}$. Finally, the resulted signal is further amplified by a factor $\eta$. Due to the fading and noisy wireless channels, the actual received signal at the aggregator is given by

$$y = \sqrt{\eta}\left(\sum_{k \in \mathcal{S}} \mathbf{m}^{\mathrm{H}}\mathbf{h}_k b_k s_k + \mathbf{m}^{\mathrm{H}}\mathbf{n}\right), \tag{4}$$

where $\mathbf{h}_k$ is the channel vector from device $k$ to the aggregator, which is assumed to be Complex Gaussian distributed with unit power, i.e., $\mathbf{h}_k \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}\right)$. This suggests that the channels are independently identically distributed (i.i.d.). Vector $\mathbf{n}$ is the additive white Gaussian noise such that $\mathcal{CN}\left(0, \sigma^2\right)$. Thus, the resulted aggregate error via AirComp is expressed as

$$e \triangleq y_{\text{des}} - y = \sum_{k \in \mathcal{S}} \left(1 - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{h}_k b_k\right) s_k - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{n}, \tag{5}$$

Letting $a_k = \left(1 - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{h}_k b_k\right)$, the MSE is then written as

$$\begin{aligned}\mathrm{MSE} = \mathbb{E}\left(\|e\|^2\right) &= \mathbb{E}\left[\left(\sum_{k\in\mathcal{S}} a_k s_k - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{n}\right)^{\mathrm{H}}\left(\sum_{k\in\mathcal{S}} a_k s_k - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{n}\right)\right] \\ &= \underbrace{\sum_{k\in\mathcal{S}} a_k^{\mathrm{H}} a_k}_{\text{fading-related error}} + \underbrace{\eta\|\mathbf{m}\|^2\sigma^2}_{\text{noise-related error}},\end{aligned} \tag{6}$$

which consists of the fading-related and the noise-related components. To reduce the error, we can eliminate the error due to the fading while mitigating the error caused by the noise. According to (6), the elimination of error caused by the fading should guarantee the following condition [24], [25]:

$$a_k = 1 - \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{h}_k b_k = 0, k \in \mathcal{S}. \tag{7}$$

Consequently, the resulted MSE are respectively expressed as

$$\mathrm{MSE} = \mathbb{E}\left(\|e\|^2\right) = \eta\|\mathbf{m}\|^2\sigma^2. \tag{8}$$

## III. Problem Formulation

When applying the AirComp technique, the signal distortion due to the fading and the noise results in the deviation of the aggregate data from the true one, which is critical for the performance of FL tasks. In the scenarios where the signal to noise ratio (SNR) is low, the resulted aggregate error under the fixed number of the selected devices can be extremely lager. Such large error degrades the learning performance [19], [26], and hence it should be reduced. On the other hand, a considerable SNR would lead to a moderate perturbation which helps improve the robustness of the neural networks [28]–[30]. Thereby, more devices need to be selected for local model update in order to enhance the learning efficiency [9], [19]. In this paper, we investigate joint device selection and aggregate beamforming design with two objectives of MSE minimization and the number of selected devices maximization. The problem formulations are presented in this section. As we have assumed before that the devices are homogeneous except for the fading channels, the device selection policy considered here is dependent on the communication factors. The consideration of learning factors such as data size, computation and storage capacities, will be left for our future works.

### *A. MSE Minimization*

To maintain the training and inference performance of the AI tasks, the large aggregate error measured by MSE should be minimized. Note that the number of the devices selected for the model is fixed, i.e., $|\mathcal{S}| = S$, where $|\mathcal{S}|$ is the cardinality of subset $\mathcal{S}$. The optimization variables include the scaling factor $\eta$, transmitting coefficient $b_k$, device subset $\mathcal{S}$, and aggregate beamforming vector $\mathbf{m}$. Since the above variables are independent from noise vector $\mathbf{n}$, minimizing the MSE in (8) shares the identical solution with the objective of $\eta \left\|\mathbf{m}\right\|^2$. Mathematically, the MSE minimization problem is formulated as

$$\min_{\mathbf{m},\eta,b_k,\mathcal{S}\subseteq\mathcal{K}} \quad \left\|\mathbf{m}\right\|^2 \eta \tag{9a}$$

$$\text{s.t.} \quad \sqrt{\eta}\mathbf{m}^{\mathrm{H}}\mathbf{h}_k b_k = 1, k \in \mathcal{S} \tag{9b}$$

$$|\mathcal{S}| = S \tag{9c}$$

$$\left\|b_k\right\|^2 \leq P, k \in \mathcal{S} \tag{9d}$$

where constraints (9b), (9c) and (9d) are respectively the condition of eliminating the fading-related error, the fixed number of the selected devices, and the transmit power constraint with $P$ the maximum power.

The optimal transmit coefficient $b_k$ following [32] can be designed as

$$b_k = \frac{\mathbf{h}_k^{\mathrm{H}}\mathbf{m}}{\sqrt{\eta}\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2}. \tag{10}$$

Substituting back to constraint (9d) yields $\eta \geq \frac{1}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2}, k \in \mathcal{S}$ for each device $k$. Thereby, $\eta$ is derived as

$$\eta = \max_k \frac{1}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2}, k \in \mathcal{S}, \tag{11}$$

which transforms problem (9) into

$$\min_{\mathbf{m},\mathcal{S}\subseteq\mathcal{K}} \quad \max_{k\in\mathcal{S}} \frac{\left\|\mathbf{m}\right\|^2}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2} \quad \text{s.t. (9c).} \tag{12}$$

Since both numerator and denominator of the above objective contain the aggregate beamforming vector $\mathbf{m}$, we can add an extra constraint such that $\left\|\mathbf{m}\right\| = 1$, which would not change the objective value. Moreover, since $\max\min 1/ = \min\max$, the above problem is equivalent to

$$\max_{\mathbf{m},\mathcal{S}\subseteq\mathcal{K}} \quad \min_{k\in\mathcal{S}} P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2 \tag{13a}$$

$$\text{s.t.} \quad \left\|\mathbf{m}\right\| = 1 \tag{13b}$$

$$\text{(9c)},$$

which is a mixed combinatorial optimization problem.

### *B. Maximizing The Number of Selected Devices*

A moderate perturbation caused by the error can be interpreted as a regulation technique, which enhances the robustness of a neural network. More devices involving in local model updating and global model aggregation can improve the learning efficiency [6] [9]. Here, we consider the maximum tolerance error denoted as $\overline{\mathrm{MSE}}$. The objective is to maximize the number of the selected devices under the MSE constraint. Same as the MSE minimization problem, the transmitting coefficient $b_k$ and the scaling factor $\eta$ are designed respectively as (10) and (11). Defining $\mathrm{MSE}_k = \frac{\|\mathbf{m}\|^2\sigma^2}{P\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\|^2}$ where $\|\mathbf{m}\| = 1$, the resulted MSE in (8) can be expressed as $\mathrm{MSE} = \max\limits_{k\in\mathcal{S}}\mathrm{MSE}_k$. Mathematically, the problem is formulated as

$$\begin{aligned}
&\max_{\mathbf{m},\mathcal{S}\subseteq\mathcal{K}} && |\mathcal{S}| && \text{(14a)}\\
&\text{s.t.} && \mathrm{MSE} = \max_{k\in\mathcal{S}}\mathrm{MSE}_k \leq \overline{\mathrm{MSE}} && \text{(14b)}\\
& && (13b),
\end{aligned}$$

which is also a mixed combinatorial optimization problem. The constraint (14b) therein indicates that the aggregate error should be less than or equal to $\overline{\mathrm{MSE}}$. communication performance.

Both considered problems are mixed combinatorial, which are difficult to solve. To obtain the optimal selected device subset, it is straight-forward to search the device subset in a brute-force manner. However, the size of the search space, i.e., $\binom{K}{S}$, becomes prohibitively large when $K$ is large, which makes brute-force searching intractable. The device subset can be determined by utilizing the sparse property as in [24], but the computational burden can be extremely large in a network with a massive number of edge devices. Moreover, optimizing the aggregate beamforming $\mathbf{m}$ for the selected devices is still challenging due to the nonconvexity. Therefore, new methods with low-complexity are needed to address these issues in a large-scale wireless network.

## IV. Random Aggregate Beamforming-Based Scheme

In this section, we propose a random aggregate beamforming-based scheme to solve the previous formulated problems, which is cost-effective. We also give the theoretical analysis in terms of the obtained MSE and the number of the selected devices. With regard to $\mathbf{m}$, we arrive at the following lemma.

**Lemma 1.** *For any given feasible* $\mathbf{m}$ *and arbitrary* $\theta \in \mathbb{R}$*, the objectives of the considered problems (13) and (14) are identical under* $\mathbf{m}$ and $\mathbf{m}e^{j\theta}$.

*Proof.* For any $\theta \in \mathbb{R}$, we have $\left\|\left(\mathbf{m}e^{j\theta}\right)^{\mathrm{H}}\mathbf{h}_k\right\|^2 = \left\|\left(\mathbf{m}\right)^{\mathrm{H}}\mathbf{h}_k\right\|^2$ and $\left\|\mathbf{m}e^{j\theta}\right\| = \|\mathbf{m}\|$, which indicate that solutions $\mathbf{m}e^{j\theta}$ and $\mathbf{m}$ have the same objective values and guarantee the constraints. □

*Remark* 1. There are infinite solutions of the aggregate beamfoming vector $\mathbf{m}$ with the same objectives. Typically, if $\mathbf{m}^*$ is the optimal solution to problems (13) and (14), $\mathbf{m}^* e^{j\theta}, \forall\theta \in \mathbb{R}$ is also optimal, which indicates that there are infinite optimal solutions of $\mathbf{m}$.

### *A. Random Aggregate Beamforming-Based Methods*

We observe that if the aggregate beamforming vector $\mathbf{m}$ is determined at the very first, the optimal devices subset for the investigated problems can be easily obtained by arranging and sorting the equivalent channel power $\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2$. Obtaining optimal $\mathcal{S}$ in this manner only requires the computational complexity of $\mathcal{O}\left(K\right)$, which is computationally efficient. Inspired by this, we propose the random aggregate beamforming-based scheme to the investigated problems. The core idea is to determine the aggregate beamforming vector $\mathbf{m}$ first, and the selected device subset $\mathcal{S}$ afterwards. Specifically, sample a vector from the set, i.e.,

$$\mathcal{M} = \left\{\mathbf{m} | \|\mathbf{m}\| {=} 1, \mathbf{m} \in \mathbb{C}^N\right\},$$

which is a complex unit sphere with $N$ dimensions. To uniformly generate $\mathbf{m}$, we can normalize a random vector $\mathbf{m}_1$ from $\mathcal{CN}\left(\mathbf{0}, \mathbf{I}\right)$, i.e., $\mathbf{m} = \mathbf{m}_1 / \left\|\mathbf{m}_1\right\|, \mathbf{m_1} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}\right)$, whose computational overhead is negligible. For problem (13), the devices subset is optimized by arranging the equivalent channel power $\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2$ and selecting the devices with the $S$ largest values. Similarly, for problem (14), the devices are selected if satisfying the condition of $\mathrm{MSE}_k \leq \overline{\mathrm{MSE}}$. It is worth noting that the proposed methods have significantly low complexity, since generating vector $\mathbf{m}$ and obtaining the subset $\mathcal{S}$ therein are easy.

The implementation of the proposed algorithms to minimize MSE and maximize the number of selected devices is shown in Fig. 2. According to this figure, the implementation of our random-based algorithms does not require the channel estimation between the aggregator and the

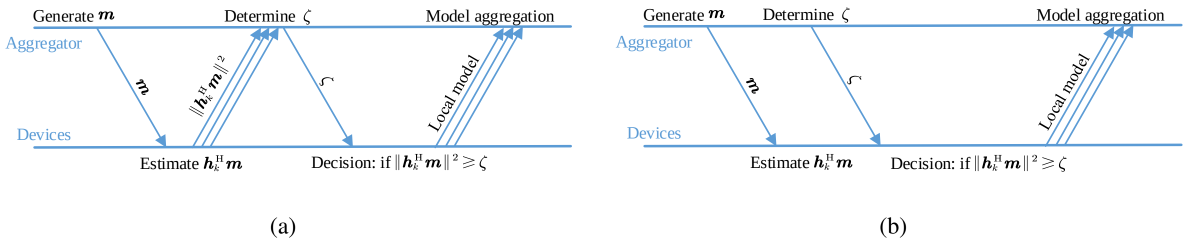

Generate $\boldsymbol{m}$
Determine $\zeta$
Model aggregation
Aggregator
Devices
$\boldsymbol{m}$
$\|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{m}\|^2$
$\zeta$
Local model
Estimate $\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{m}$
Decision: if $\|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{m}\|^2 \geqslant \zeta$
(a)
Generate $\boldsymbol{m}$
Determine $\zeta$
Model aggregation
Aggregator
Devices
$\boldsymbol{m}$
$\zeta$
Local model
Estimate $\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{m}$
Decision: if $\|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{m}\|^2 \geqslant \zeta$
(b)

Figure 2: Implementation of the proposed scheme for (a) problem (13) and (b) problem (14).

edge devices, which can greatly reduce the implementation complexity. For MSE minimization problem (13), the aggregator first generates a random aggregate beamforming vector $\mathbf{m}$, which is broadcast to all devices. Then all devices estimate the obtained value $\mathbf{h}_k^{\mathrm{H}}\mathbf{m}$ which is a weighted channel value rather than $\mathbf{h}_k^{\mathrm{H}}$, and feedback $\|\mathbf{h}_k^{\mathrm{H}}\mathbf{m}\|^2$ to the aggregator. After receiving the feedbacks from all devices, the aggregator determines and broadcasts a threshold $\zeta$, which is obtained by finding the $S$ largest weighted channel gain values. Finally, the devices are selected for local model updates and global model aggregation if their weighted channel gain values are no less than $\zeta$. For problem (14), the implementation procedure is similar except for the determination of $\zeta$. The threshold therein is determined directly at the aggregator without the information of weighted channel gains, which is obtained such that $\zeta = \frac{\sigma^2}{P\mathrm{MSE}}$. The implementation of our proposed algorithms does not require any channel information. This is the major advantage compared with those methods requiring the channel information, since the overhead for channel estimation in a lager-scale network is heavy.

In the following, we focus on the theoretical analysis in terms of the objectives of MSE and the number of the selected devices in a large-scale system with massive edge devices.

### *B. The Distribution of* $\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|$

We first analyze the property of inner product $\mathbf{m}^{\mathrm{H}}\mathbf{h}_k = \sum_{i=1}^{N}\mathbf{m}[i]\mathbf{h}_k[i]$ where $\mathbf{m}[i]$ and $\mathbf{h}_k[i]$ are the $i$-th element of vector $\mathbf{m}$ and $\mathbf{h}_k$. Since $\mathbf{h}_k[i]$ is Gaussian distributed, i.e., $\mathbf{h}_k[i] \sim \mathcal{CN}(0,1)$, we have $\mathbf{m}[i]\mathbf{h}_k[i] \sim \mathcal{CN}(0, \|\mathbf{m}[i]\|^2)$. Together with the additional constraint of $\|\mathbf{m}\| = 1$, the distribution of $\mathbf{m}^{\mathrm{H}}\mathbf{h}_k$ is readily obtained as

$$\mathbf{m}^{\mathrm{H}}\mathbf{h}_k \sim \mathcal{CN}(0, \sum_{i=1}^{N}\|\mathbf{m}[i]\|^2) = \mathcal{CN}(0,1), \tag{15}$$

which is Gaussian distributed with mean $0$ and variance $1$. Since the channels $\mathbf{h}_k, \forall k \in \mathcal{K}$ are i.i.d., the new random variables $\mathbf{m}^{\mathrm{H}}\mathbf{h}_k, k \in \mathcal{K}$ are i.i.d.. Their modulus $\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|, \forall k \in \mathcal{K}$

are also i.i.d., which are Rayleigh distributed. By defining a new variable $Z_k = \left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|$, the probability density function (PDF) is expressed as

$$f(z_k) = 2z_k e^{-z_k^2}, z_k \geq 0, \tag{16}$$

whose cumulative density probability (CDF) is given by

$$\mathbb{P}\left(Z_k < z_k\right) = 1 - e^{-z_k^2}, z_k \geq 0. \tag{17}$$

## *C. MSE Minimization*

For the problem of MSE minimization, the device subset is obtained by arranging the equivalent channel gain based on the sampled aggregate beamforming vector from a complex unit sphere. In this case, the obtained MSE performance given $\mathbf{m}$ is written as

$$\frac{\mathrm{MSE}}{\sigma^2} = \min_{\mathcal{S}\subseteq\mathcal{K}} \max_{k\in\mathcal{S}} \frac{1}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2}. \tag{18}$$

Defining a new random variable $Y_k = \left(P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2\right)^{-1}$, we then have $Y_k = \left(PZ_k^2\right)^{-1}$, which is i.i.d.. According to (17), the CDF of $Y_k$ is obtained as

$$F(y_k) = \mathbb{P}\left(Y_k < y_k\right) = \mathbb{P}\left(Z_k > \left(\sqrt{Py_k}\right)^{-1}\right) = e^{-(Py_k)^{-1}}, y_k > 0. \tag{19}$$

Thereby, the PDF of variable $Y_k$ is obtained by deriving the derivative of (19), given by

$$f\left(y_k\right) = \left(Py_k^2\right)^{-1} e^{-(Py_k)^{-1}}, y_k > 0. \tag{20}$$

As mentioned, variables $Y_1, Y_2, \ldots, Y_K$ for all devices are independent with the same distribution, which can be considered as a sequence of variables sampled from the absolutely continuous population with PDF $f(y)$ in (20) and CDF $F\left(y\right)$ in (19).

Define another variable $X$ as the objective value of problem (13), i.e.,

$$X = \min_{\mathcal{S}\subseteq\mathcal{K}} \max_{k\in\mathcal{S}} \frac{1}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2} = \min_{\mathcal{S}\subseteq\mathcal{K}} \max_{k\in\mathcal{S}} Y_k. \tag{21}$$

Let $Y_{1:K} \leq Y_{2:K} \leq \cdots \leq Y_{K:K}$ be the order statistics obtained by arranging the variables $Y_1, Y_2, \ldots, Y_K$ in an ascending order. We then have $X = Y_{S:K}$. The CDF of $X$ can be derived without much difficulty such that

$$\begin{aligned} G(x) &= \mathbb{P}(X < x) = \mathbb{P}(Y_{S:K} < x) = 1 - \mathbb{P}(Y_{S:K} > x) \\ &= 1 - \mathbb{P}(\text{at most } S-1 \text{ of } Y_1, Y_2, \ldots, Y_K \text{ are at most } x) \\ &= 1 - \sum_{s=0}^{S-1} \mathbb{P}(\text{exactly } s \text{ of } Y_1, Y_2, \ldots, Y_K \text{ are at most } x) \\ &= 1 - \sum_{s=0}^{S-1} \binom{K}{s} \left[\mathbb{P}(Y < x)\right]^s \left[\mathbb{P}(Y > x)\right]^{K-s} \\ &= 1 - \sum_{s=0}^{S-1} \binom{K}{s} e^{-s(Px)^{-1}} \left[1 - e^{-(Px)^{-1}}\right]^{K-s} \end{aligned} \tag{22}$$

In a large-scale distributed system, there are a large population of devices involving in the implementation of edge intelligence. Thereby, it is also important to analyze the obtained MSE performance when $K$ is large. The following lemma establishes the asymptotic distribution of a central order statistic when $K \to \infty$.

**Lemma 2.** *Denoting $q = S/K$, as $K \to \infty$, we have*

$$\sqrt{K} f\left(F^{-1}(q)\right) \frac{\left(X - F^{-1}(q)\right)}{\sqrt{q(1-q)}} \xrightarrow{d} \mathcal{N}(0,1),$$

*where $F^{-1}(\cdot)$ is the inverse function of CDF $F$ in (19), $\xrightarrow{d}$ means convergence in distribution.*

*Proof.* According to (19) and (20), both CDF $F$ and PDF $f$ are continuous in their domains. Since $0 < S < K$, we have $q \in (0,1)$ and then have $F^{-1}(q) = -\frac{1}{P \ln q} > 0$. Therefore, the condition for Theorem 8.5.1 in [33] is satisfied[1]. This completes the proof. □

Lemma 2 shows that random variable $X$ is asymptotically normal after suitable normalization. It also indicates that the expectation and the variance of MSE performance $X$ can be approximated with probability $q = S/K$, expressed by

$$\mathbb{E}(X) \simeq F^{-1}(q), \tag{23}$$

[1] Let $X_{i:n}$ be the $i$-th order statistic from $n$ random variables. The theorem is elaborated such that for $0 < p < 1$, let $F$ be absolutely continuous with PDF $f$ which is positive at $F^{-1}(p)$ and is continuous at that point. For $i = \lfloor np \rfloor + 1$, as $n \to \infty$, we have $\sqrt{n} f\left(F^{-1}(p)\right) \frac{\left(X_{i:n} - F^{-1}(p)\right)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$.

$$\mathrm{VAR}\left(X\right) \simeq \frac{q\left(1-q\right)}{K\left[f\left(F^{-1}\left(q\right)\right)\right]^2} = \frac{\left(1-q\right)}{KP^2 q\left(\ln q\right)^4}, \tag{24}$$

where $F^{-1}\left(q\right)$ and $f\left(F^{-1}\left(q\right)\right)$ are obtained from (19) and (20) such that

$$F^{-1}\left(q\right) = -\frac{1}{P\ln q},\ f\left(F^{-1}\left(q\right)\right) = Pq\left(\ln q\right)^2.$$

**Theorem 1.** *The proposed random aggregate beamforming-based method for problem (13) approaches the optimal solution when $K \to \infty$ and $S \ll K$.*

*Proof.* For the number of devices $K \to \infty$ and a fixed number $S \ll K$ , we have $q = S/K \to 0$. The limit of the right terms of (23) and (24) are respectively given by

$$\lim_{q\to 0} F^{-1}\left(q\right) = \lim_{q\to 0} -\frac{1}{P\ln q} = 0. \tag{25}$$

Thus, we have $\mathbb{E}\left(X\right) \to 0$ for $K \to \infty$ and $S \ll K$.

Denoting $X^*$ as the optimal objective of problem (12), we have $X \geq X^* > 0$. Then the expectations of $X$ and $X^*$ satisfy:

$$\mathbb{E}\left(X\right) \geq \mathbb{E}\left(X^*\right) > 0. \tag{26}$$

From the above observations, we have $\mathbb{E}\left(X^*\right) \to 0$ and further arrive at $\mathbb{E}\left(X\right) = \mathbb{E}\left(X^*\right)$, which indicates that the proposed random aggregate beamforming-based method approaches the optimal solution for $K \to \infty, S \ll K$. □

### *D. Maximum Number of Selected Devices*

Similar to the previous subsection, the solution of the aggregate beamforming vector $\mathbf{m}$ to problem (14) is determined by randomly sampling a vector from $\mathcal{M}$. Then the devices that guarantee the MSE constraint are selected, i.e.,

$$\begin{aligned}\mathcal{S} &= \left\{k|\mathrm{MSE}_k \leq \overline{\mathrm{MSE}}, \forall k \in \mathcal{K}\right\} \\ &= \left\{k|\frac{1}{P\left\|\mathbf{m}^{\mathrm{H}}\mathbf{h}_k\right\|^2} \leq \frac{\overline{\mathrm{MSE}}}{\sigma^2}, \forall k \in \mathcal{K}\right\} \\ &= \left\{k|Y_k \leq \tilde{x}, \forall k \in \mathcal{K}\right\}, \end{aligned} \tag{27}$$

where $\tilde{x} = \overline{\mathrm{MSE}}/\sigma^2$. Defining the discrete random variable $\Lambda = |\mathcal{S}|$, the probability mass function (PMF) is obtained as

$$\begin{aligned}\mathbb{P}\left(\varLambda = S\right) &= \mathbb{P}\left(\text{Exactly } S \text{ of } Y_1, Y_2, \ldots, Y_K \text{ are less than } \tilde{x}\right)\\ &= \binom{K}{S}\prod_{i=1}^{S}\mathbb{P}\left(Y < \tilde{x}\right)\prod_{i=1}^{K-S}\mathbb{P}\left(Y > \tilde{x}\right)\\ &= \binom{K}{S}e^{-S(P\tilde{x})^{-1}}\left(1 - e^{-(P\tilde{x})^{-1}}\right)^{K-S}.\end{aligned} \tag{28}$$

In a large-scale system where there are massive devices, we arrive at the following theorem.

**Theorem 2.** *Denote* $\Lambda_{\min} = \frac{K}{e^{1/(P(\tilde{x}-3\sigma))}}$, $\Lambda_{\max} = \frac{K}{e^{1/(P(\tilde{x}+3\sigma))}}$, *where* $\sigma = \sqrt{\frac{1-\exp\left(-\frac{1}{P\tilde{x}}\right)}{KP^2\exp\left(-\frac{1}{P\tilde{x}}\right)\left(-\frac{1}{P\tilde{x}}\right)^4}}$. *The random variable* $\Lambda$ *is approximately symmetric within the interval of* $[\Lambda_{\min}, \Lambda_{\max}]$ *with the expectation of* $\mathbb{E}\left(\Lambda\right) = Ke^{-1/(P\tilde{x})}$ *when* $K \to \infty$.

*Proof.* When the expectation of the MSE performance in the MSE minimization problem equals to $\tilde{x}$ for $K \to \infty$, we have

$$\begin{aligned}\mathbb{E}\left(X\right) \simeq & F^{-1}\left(q\right) = -\frac{1}{P\ln q} = \tilde{x}\\ \Leftrightarrow & -\frac{1}{P\ln q} = \tilde{x} \Leftrightarrow q = \exp\left(-\frac{1}{P\tilde{x}}\right).\end{aligned}$$

Then the variance of $X$ can be expressed as

$$\mathrm{VAR}\left(X\right) \simeq \frac{q\left(1-q\right)}{K\left[f\left(F^{-1}\left(q\right)\right)\right]^2} = \frac{1-\exp\left(-\frac{1}{P\tilde{x}}\right)}{KP^2\exp\left(-\frac{1}{P\tilde{x}}\right)\left(-\frac{1}{P\tilde{x}}\right)^4}, \tag{29}$$

which approaches $0$ when $K \to \infty$. Since the random variable $X$ is asymptotically normal, variable $X$ is within the interval of $[\tilde{x} - 3\sigma, \tilde{x} + 3\sigma]$ where $\sigma = \sqrt{\frac{1-\exp\left(-\frac{1}{P\tilde{x}}\right)}{KP^2\exp\left(-\frac{1}{P\tilde{x}}\right)\left(-\frac{1}{P\tilde{x}}\right)^4}}$ is the standard deviation.

For any MSE performance variable $X'$ whose expectation $\mathbb{E}\left(X'\right)$ is within the interval of $\in [\tilde{x} - 3\sigma, \tilde{x} + 3\sigma]$, its variance $\mathrm{VAR}\left(X'\right)$ approximately equals $\mathrm{VAR}\left(X\right)$ owing to the small standard deviation $\sigma$. Similarly, the probability $q'$ for variable $X'$ is written as $q = \exp\left(-\frac{1}{P\mathbb{E}(X')}\right)$. Thus, the corresponding number of the selected devices is $S = \left\lfloor\frac{K}{e^{1/(P\mathbb{E}(X'))}}\right\rfloor$. In the following, we consider two cases where $\mathbb{E}\left(X'\right)$ is not in the interval of $[\tilde{x} - 3\sigma, \tilde{x} + 3\sigma]$.

Case1: When $\mathbb{E}\left(X'\right) < \tilde{x} - 3\sigma$, we have $S < \frac{K}{e^{1/(P(\tilde{x}-3\sigma))}}$ and $X'$ is approximately within $[\mathbb{E}\left(X'\right) - 3\sigma, \mathbb{E}\left(X'\right) + 3\sigma]$. Since $\mathbb{E}\left(X'\right) + 3\sigma < \tilde{x}$, it shows that selecting $S$ devices guarantees the MSE constraint, which further suggests that more devices can be selected.

**Algorithm 1** Random Aggregate Beamforming-based Design for MSE Minimization.

**Input**: $N_m$, $S$ **Output**: $\mathcal{S}$, $\mathbf{m}$ **Initialize**: $\mathrm{val} = 0$, $\mathrm{tmp}_{\max} = 0$

```
for $n = 1 : N_m$
    Sample $\mathbf{m}_r$ from $\mathcal{CN}(0, \mathbf{I})$
    Normalize $\mathbf{m}_r = \mathbf{m}_r / \|\mathbf{m}_r\|$
    Calculate $\mathrm{tmp_k} = \left\|\mathbf{m}_r^{\mathrm{H}}\mathbf{h}_k\right\|^2, \forall k \in \mathcal{K}$
    Find the $S$-th largest $\mathrm{tmp_k}, \forall k \in \mathcal{K}$
    If $\mathrm{tmp} > \mathrm{tmp}_{\max}$.
        $\mathcal{S} = \left\{k \mid \left\|\mathbf{m}_r^{\mathrm{H}}\mathbf{h}_k\right\|^2 \geq \mathrm{tmp}_{\max}, \forall k \in \mathcal{K}\right\}$
        $\mathbf{m} = \mathbf{m}_r$.
```

Case2: When $\mathbb{E}\left(X'\right) > \tilde{x} + 3\sigma$, we have $S > \frac{K}{e^{1/(P(\tilde{x}+3\sigma))}}$ and $X'$ is approximately within $\left[\mathbb{E}\left(X'\right) - 3\sigma, \mathbb{E}\left(X'\right) + 3\sigma\right]$. Since $\mathbb{E}\left(X'\right) - 3\sigma > \tilde{x}$, the MSE constraint cannot be guaranteed if selecting $S > \frac{K}{e^{1/(P(\tilde{x}+3\sigma))}}$ devices.

Thus, the approximate minimum and maximum number of the selected devices under the random aggregate beamforming based method are respectively

$$\Lambda_{\min} = \frac{K}{e^{1/(P(\tilde{x}-3\sigma))}}, \quad \Lambda_{\max} = \frac{K}{e^{1/(P(\tilde{x}+3\sigma))}}. \tag{30}$$

As stated above, when the number of the selected devices $S \in [\Lambda_{\min}, \Lambda_{\max}]$, the obtained MSE performance has almost the same variance. Therefore, the PMF of variable $\Lambda$ is symmetric within the interval of $[\Lambda_{\min}, \Lambda_{\max}]$ with the expectation of $\mathbb{E}\left(\Lambda\right) = Ke^{-1/(P\tilde{x})}$. This completes the proof. □

## V. Refined Method

The previous section first gives the proposed random aggregate beamforming-based scheme, and then presents theoretical analysis in terms of the two objectives when $K$ is large. This section focuses on the performance improvement of the proposed methods when $K \ll \infty$. Specifically, we refine the proposed scheme by randomizing the vector $\mathbf{m}$ from $\mathcal{M}$ for $N_m$ times. The solutions with the best performance are obtained among $N_m$ aggregate beamforming vectors. The refined methods for both MSE minimization and the number of selected devices maximization are detailed in Algorithm 1 and Algorithm 2. Additionally, the effectiveness of the refined methods are analyzed.

Denote $\mathbf{m}^*$ and $\mathcal{S}^*$ as the the optimal aggregate beamforming vector and selected device subset, respectively. According to Lemma 1, there are infinite optimal solutions, i.e., $\mathbf{m}^* e^{j\theta}, \forall \theta \in \mathcal{R}$,

**Algorithm 2** Random Aggregate Beamforming-based Design for the Number of Selected Devices Minimization.

**Input**:$N_m$, $\overline{\mathrm{MSE}}/\sigma^2$ **Output**: $\mathcal{S}_{\max}$, $\mathbf{m}$ **Initialize**: $\mathrm{S}_{\max} = 0$

```
for $n = 1 : N_m$
    Sample $\mathbf{m}_r$ from $\mathcal{CN}(0, \mathbf{I})$
    Normalize $\mathbf{m}_r = \mathbf{m}_r / \left\| \mathbf{m}_r \right\|$
    Calculate $\mathrm{tmp}_k = \left( P \left\| \mathbf{m}_r^{\mathrm{H}} \mathbf{h}_k \right\|^2 \right)^{-1}, \forall k \in \mathcal{K}$
    $\mathcal{S} = \left\{ k \left| \mathrm{tmp}_{\mathrm{k}} \leq \overline{\mathrm{MSE}}/\sigma^2, \forall k \in \mathcal{K} \right. \right\}$
    If $|\mathcal{S}| > |\mathcal{S}_{\max}|$
        $\mathcal{S}_{\max} = \mathcal{S}$
        $\mathbf{m} = \mathbf{m}_r$.
```

with the same optimal objectives of the MSE and the number of selected devices, which are denoted as $x^*$ and $\gamma^*$ respectively. As the aggregate beamforming vector $\mathbf{m}$ is uniformly sampled from the complex unit sphere $\mathcal{M}$, the probability that the sampled vector $\mathbf{m}$ is optimal, i.e., $\mathbb{P}(\mathbf{m} = \mathbf{m}^*)$, can be interpreted as the proportion of the region of the optimal solutions on the whole sphere $\mathcal{M}$. Despite infinite optimal solutions of $\mathbf{m}$, the probability can be very small. Nevertheless, we can model the probability of a sub-optimal solutions of both problems in an implicit manner, which are detailed in the next subsections.

### *A. MSE Minimization*

For problem (13), the optimal objective as $x^*$ is related to the channel of all devices $\mathbf{h}_k, \forall k \in \mathcal{K}$. Let $\mathit{\Delta} \in \mathbb{R}^+$ be the difference between the optimal objective and the sub-optimal one. Then, the probability that the obtained sub-optimal objective under the proposed method less than a given value $x$ where $x > x^*$ can be modelled as

$$\varsigma = \mathbb{P}\left(X < x; \{\mathbf{h}_k, \forall k \in \mathcal{K}\}\right) = \mathbb{P}\left(X < x^* + \mathit{\Delta}\right). \tag{31}$$

The optimal objective $x^*$ is difficult to obtain as the problem (13) is non-convex. In addition, it is difficult to obtain the distribution of objective $X$ under $\mathbf{h}_k, \forall k \in \mathcal{K}$, where $\mathbf{m}$ is uniformly distributed on the unit sphere. Thus, the probability $\varsigma$ cannot be analytically expressed. Instead, we model $\sigma$ as an implicit function of $x^*$ and $\mathit{\Delta}$, i.e., $\varsigma = H\left(\mathit{\Delta}; x^*\right)$ with the range of $[0, 1]$. The function is monotonically nondecreasing w.r.t. the difference $\mathit{\Delta} > 0$, which is useful for our later analysis. To illustrate the distribution, we search the feasible solutions to obtain $H\left(\mathit{\Delta}; x^*\right)$

under one channel realization in the scenario with $K = 50$ devices and an aggregator equipped with $N = 2$ antennas. As can be seen in Fig. 3, for the optimal objective $x^* = 0.455$, the probability $\varsigma$ is 0.01 when the difference $\mathit{\Delta} = 0.011$. By randomizing the vector $\mathbf{m}$ from $\mathcal{M} = \{\mathbf{m}|\|\mathbf{m}\|{=}1, \mathbf{m} \in \mathbb{C}^2\}$ for $100$ times, we get $63.4\%$ chance that the obtanied objective $x$ is whithin the interval of $[0.455, 0.466]$.

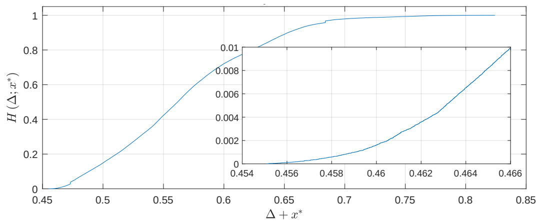


Figure 3: Distribution $H\left(\mathit{\Delta}; x^*\right)$ under one channel realization with $S = 10$ devices are selected.

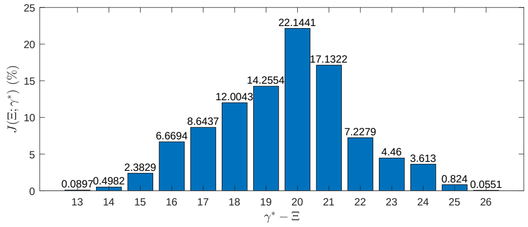


Figure 4: PMF function $\beta = J\left(\Xi; \gamma^*\right)$ under one channel realization with the threshold $\overline{\mathrm{MSE}}/\sigma^2 = 1$.

Considering $N_m$ randomizations of the aggregate vectors, we arrive at the following theorem.

**Theorem 3.** *Let* $\bar{X} = \min\left(X_1, \ldots, X_i, \ldots, X_{N_m}\right)$ *be the obtained objective value given* $N_m$ *random aggregate beamforming vectors under the channel conditions of* $\mathbf{h}_k, \forall k \in \mathcal{K}$. *For any objective value difference* $\mathit{\Delta} > 0$, *we have*

$$\lim_{N_m \to \infty} \mathbb{P}\left(\bar{X} < x^* + \mathit{\Delta}\right) = 1.$$

*Proof.* Since $N_m$ random aggregate beamforming vectors are uniformly sampled from the same complex unit sphere $\mathcal{M}$, the probability of $\bar{X}$ to be less than $x = x^* + \mathit{\Delta}$ is given by

$$\begin{aligned}
\mathbb{P}\left(\bar{X} < x\right) &= \mathbb{P}\left(\bar{X} < x^* + \mathit{\Delta}\right) \\
&= 1 - \mathbb{P}\left(\bar{X} > x^* + \mathit{\Delta}\right) \\
&= 1 - \prod_{i=1}^{N_m} \mathbb{P}\left(X_i > x^* + \mathit{\Delta}\right) \\
&= 1 - \left(1 - H\left(\mathit{\Delta}; x^*\right)\right)^{N_m},
\end{aligned}$$

where $H\left(\cdot\right)$ is a monotonic function w.r.t the difference $\mathit{\Delta}$. Given $\mathit{\Delta} > 0$, for any $\delta > 0$, the following equivalence holds, i.e.,

$$\begin{aligned}&\left|\mathbb{P}\left(\bar{X} < x^* + \mathit{\Delta}\right) - 1\right| < \delta \\ \Leftrightarrow &\left(1 - H\left(\mathit{\Delta}; x^*\right)\right)^{N_m} < \delta \\ \Leftrightarrow &N_m > \left\lfloor \log_{1-H(\mathit{\Delta};x^*)} \delta \right\rfloor ,\end{aligned}$$

where $\lfloor\cdot\rfloor$ is the floor function. Therefore, given $\mathit{\Delta} > 0$, for $\forall\delta > 0$, there exist a number $N = \left\lfloor \log_{1-H(\mathit{\Delta};x^*)} \delta \right\rfloor$ such that when $N_m > N$, we have $\left|\mathbb{P}\left(\bar{x} < x^* + \mathit{\Delta}\right) - 1\right| < \delta$. This completes the proof. □

### *B. Maximizing The Number of Selected Devices*

As for the mixed combinatorial problem (14), the optimal objective $\gamma^*$ is difficult to obtain, which does not have the close-form expression w.r.t. $\mathbf{h}_k, \forall k \in \mathcal{K}$. Also, the PMF of the number of selected devices $\Lambda$ obtained by the proposed method is unknown under $\mathbf{h}_k, \forall k \in \mathcal{K}$. Thus, the probability that the number of selected devices is greater than a sub-optimal objection $\gamma$ cannot be explicitly expressed, i.e.,

$$\beta = \mathbb{P}\left(\Lambda = \gamma\right) = \mathbb{P}\left(\Lambda = \gamma^* - \Xi\right), \tag{32}$$

where $\Xi \in \mathbb{N}$ is the difference between the optimal number of selected devices and the sub-optimal one. It is a function of $\gamma^*$ and $\Xi$, i.e., $\beta = J\left(\Xi; \gamma^*\right)$, which is monotonically nondecreasing w.r.t. $\Xi$ within the range of $[0, 1]$. Likewise, we illustrate the probability $J\left(\Xi; \gamma^*\right)$ by searching the feasible solutions under one channel realization in a scenario with $K = 50$ devices and an aggregator equipped with $N = 2$ antennas. As can be seen in Fig. 4, the optimal objective $\gamma^* = 26$, the probability of which is $0.0551\%$. In this case, $1000$ times of randomization can lead to $42.37\%$ chance to obtain the optimal solution.

To improve the performance, $N_m$ randomizations of the aggregate vectors are considered. The following Theorem can be arrived.

**Theorem 4.** *Let $\bar{\mathit{\Lambda}} = \min\left(\mathit{\Lambda}_1, \cdots, \mathit{\Lambda}_i, \cdots, \mathit{\Lambda}_{N_m}\right)$ be the obtained objective value via the refined method, where $\mathit{\Lambda}_i$ is the objective value under the $i$-th random vector $\mathbf{m}_i$. For any objective value difference $\Xi \in \mathbb{N}$, we have*

$$\lim_{N_m\to\infty} \mathbb{P}\left(\bar{\mathit{\Lambda}} > \gamma^* - \Xi\right) = 1.$$

*Proof.* The proof is similar to that of Theorem 3. The probability of $\bar{\mathit{\Lambda}}$ to be greater than $\gamma = \gamma^* - \Xi$ is implicitly expressed as $\mathbb{P}\left(\bar{\mathit{\Lambda}} > \gamma\right) = \mathbb{P}\left(\bar{\mathit{\Lambda}} > \gamma^* - \Xi\right) = 1 - \left(1 - J\left(\Xi; \gamma^*\right)\right)^{N_m}$. Given $\Xi \in \mathbb{N}$, for $\forall \delta > 0$, there exists the number $N = \left\lfloor \log_{1-J(\Xi;\gamma^*)} \delta \right\rfloor$ such that when $N_m > N$, we have $\left|\mathbb{P}\left(\bar{\mathit{\Lambda}} > \gamma^* - \Xi\right) - 1\right| < \delta$. This completes the proof. □

### C. Observations

When $K \ll \infty$, we refine the proposed methods in Section IV by randomizing $N_m$ aggregate beamforming vectors aiming at the performance improvement. The proposed methods in Section IV have a computational complexity of $\mathcal{O}\left(K\right)$, and thereby the required computational complexity for the refined methods is $\mathcal{O}\left(N_m K\right)$. From the above analysis, the gaps between the suboptimal performance and the optimal ones are less than $\mathit{\Delta}$ and $\Xi$ when $N_m > \left\lfloor \log_{1-H(\mathit{\Delta};x^*)} \delta \right\rfloor$ and $N_m > \left\lfloor \log_{1-J(\Xi;\gamma^*)} \delta \right\rfloor$ where $\delta > 0$ is arbitrary small. Reversely, given the randomization number $N_m$, the corresponding differences are respectively $\mathit{\Delta} < H^{-1}\left(1 - \sqrt[N_m]{\delta}; x^*\right)$ and $\Xi < J^{-1}\left(1 - \sqrt[N_m]{\delta}; \gamma^*\right)$. Despite that the inverse functions $H^{-1}\left(\cdot\right)$ and $J^{-1}\left(\cdot\right)$ cannot be explicitly expressed, the refined methods with multiple randomizations provide an obvious insight of performance improvement. The reason behind is that both functions $H$ and $J$ are monotonically nondecreasing w.r.t. $\mathit{\Delta}$ and $\Xi$, which indicate that the inverse functions are also monotonically nondecreasing w.r.t. $1 - \sqrt[N_m]{\delta}$. Therefore, a larger randomlization $N_m$ leads to a small differences $\mathit{\Delta}$ and $\Xi$ from the optimal objectives.

## VI. Simulation

This section presents the simulation results to demonstrate the effectiveness of the proposed random aggregate beamforming-based methods as well as the refined ones. Besides, the theoretical analysis is confirmed. The maximum transmit power of the devices is set to $P = 0$ dB. To showcase the advantage of the proposed method, we compare it to reference methods in terms of the MSE performance and the number of the selected devices for the investigated problems. For the problem of MSE minimization, the benchmarks are

- Iterative device selection and aggregate beamforming design: Given the selected devices, the aggregate beamforming vector is obtained by utilizing DC method. After the aggregate beamforming vector is obtained, arrange the equivalent channel power and select the devices with $S$ largest values. The solutions are obtained by alternatively solving the above two subproblems until convergence.

- Random device selection and aggregate beamforming design: Randomly select $S$ out of $K$ devices, and optimize the aggregate beamforming vector by DC technique.

The problem of the number of selected devices maximization can be modelled as a sparse problem. DC representation and $\ell_1-\mathrm{norm}$ techniques are the two most common methods to handle the sparsity. The comparing benchmarks are:

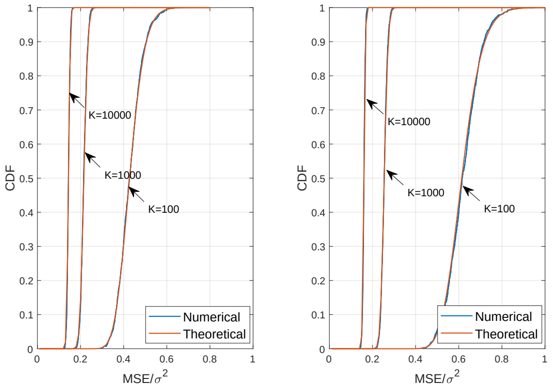
- A novel DC algorithm [24]: The DC representation is adopted to induce the sparsity, based on which the DC technique is used to obtain the aggregate beamforming vector.
- $\ell_1$+SDR [34]: The sparsity is induced by the $\ell_1-\mathrm{norm}$ technique, and the aggregate beamforming vector is obtained by SDR method.

### A. Analysis Verification

To verify the theoretical analysis in terms of the MSE and the selected number under the proposed random aggregate beamforming vector, we conduct 1000 Monte Carlo simulations, where the number of antennas at the aggregator is set to $N=8$.

For the problem of MSE minimization, we consider two settings where the fixed numbers of selected devices $S$ are $10$ and $20$ respectively. Fig. 5 displays the theoretical and empirical CDFs under these two settings, where the coinciding lines validate the derived distribution of $\mathrm{MSE}/\sigma^2$ in (22). By comparing Fig. 5(a) and Fig. 5(b), we can observe that selecting more devices leads to a larger aggregate error, because the scaling factor $\eta$ shall be increased in order to align signals from more devices. In addition, as the increase of the number of devices $K$, the obtained $\mathrm{MSE}/\sigma^2$ together with its variance decrease with the trends towards $0$. To confirm the Lemma 2, we compare the CDF of normal distribution with mean $F^{-1}\left(q\right)$ and variance $\frac{(1-q)}{KP^2q(\ln q)^4}$ with the numerical one, which is displayed in Fig. 6. It shows that the distribution of the obtained $\mathrm{MSE}/\sigma^2$ converges to the normal distribution as the increase of $K$, which confirms Lemma 2. Moreover, the mean and variance of $\mathrm{MSE}/\sigma^2$ are shown to approach $0$ when $K$ is getting larger, which validates Theorem 1.

For the problem of the maximum number of selected devices, we present both the numerical PMF and the the derived one, where the MSE threshold is set to $\overline{\mathrm{MSE}}/\sigma^2=0.2$. As is shown in Fig. 7, the numerical PMF has approximately the same number of the selected devices as the derived one, which ensures the correctness of (7). In addition, the PMF becomes symmetric with the increase of $K$. To verify Theorem 2, we calculate the average, minimum and maximum number of selected devices in the scenarios with the threshold $\overline{\mathrm{MSE}}/\sigma^2$ set to $0.2$, $0.35$ and $0.5$.



(a)

(b)

Figure 5: Derived and numerical CDFs. (a) $S = 10$. (b) $S = 20$.

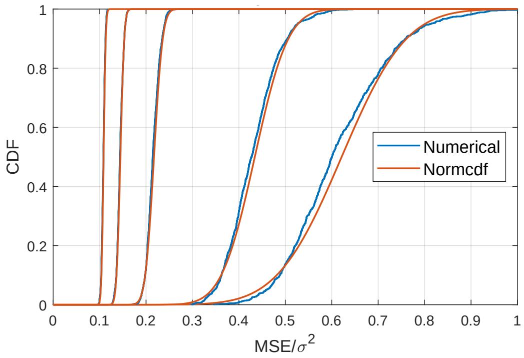
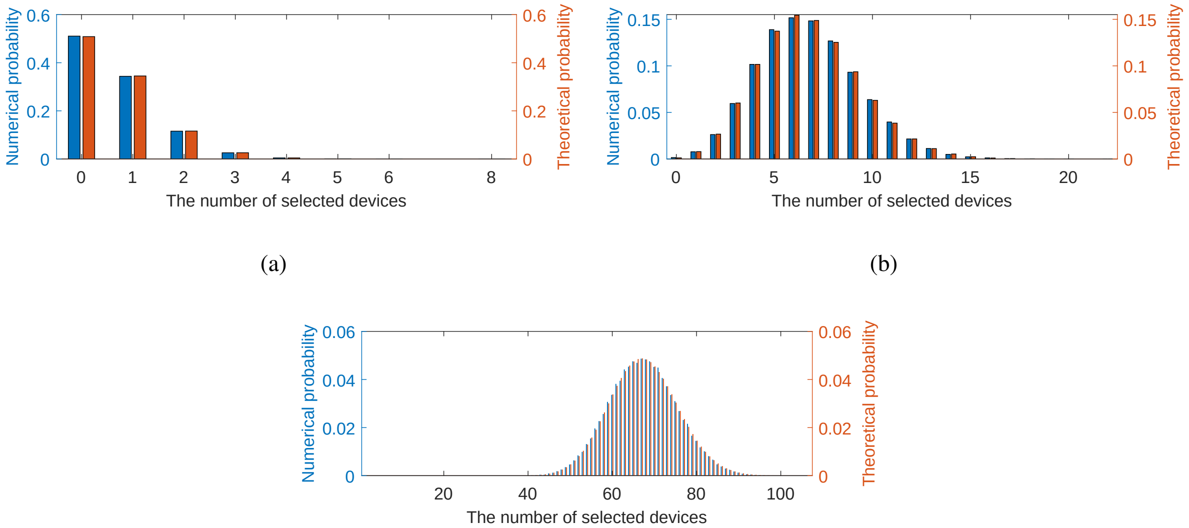


Figure 6: Empirical CDF and the approximated normal distribution. The number of the devices from right to left are respectively: $K = 50, K = 10^2, K = 10^3, K = 10^4, K = 10^5$.



(a)

(b)

(c)

Figure 7: Numerical and theoretical PMFs. (a) $K = 10^2$. (b) $K = 10^3$. (c) $K = 10^4$.

Table I gives the average, minimum and maximum number of the selected devices. From the Table, we can see that the average number of the selected devices are almost the same for both numerical and theoretical results. The numerical minimum number is smaller than the theoretical one, while the numerical maximum number is larger than the theoretical one. These relative gaps are diminished as the increase of $K$, which verifies Theorem 2.

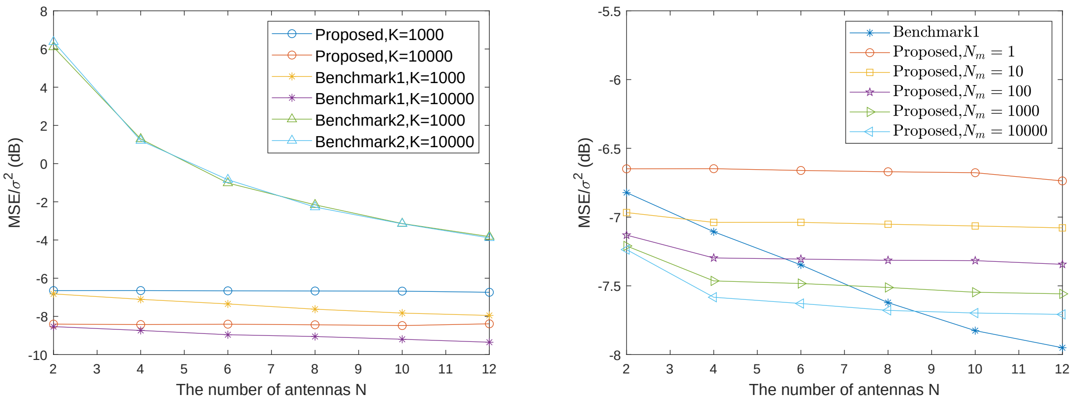
| Scenario | Threshold $\overline{\text{MSE}}/\sigma^2$ | 0.2 | | | 0.35 | | | 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Stochastic | Average | Minimum | Maximum | Average | Minimum | Maximum | Average | Minimum | Maximum |
| $K = 10^4$ | Numerical | 67.36 | 37 | 103 | 574.21 | 479 | 656 | 1353.49 | 1234 | 1487 |
| | Theoretical | 67.38 | 52.20 | 84.94 | 574.33 | 528.38 | 621.41 | 1353.35 | 1284.95 | 1421.76 |
| $K = 10^5$ | Numerical | 674.45 | 585 | 778 | 5744.21 | 5546 | 6046 | 13533.16 | 13133 | 13908 |
| | Theoretical | 673.79 | 598.86 | 754.09 | 5743.26 | 5596.68 | 5890.97 | 13533.53 | 13317.18 | 13749.87 |

Table I: Average, minimum and maximum values of the number of the selected devices.

### *B. MSE Performance*

In order to showcase the performance of the proposed random aggregate beamforming-based design, we conduct simulations for 100 channel realizations. The proposed methods are compared to the benchmarks previously detailed in the scenarios with different number of antennas $N$ at the aggregator. We label the iterative device selection and aggregate beamforming design and the random device selection and aggregate beamforming design as 'Benchmark1' and 'Benchmark2' respectively. As shown in Fig. 8(a), MSE performance $\text{MSE}/\sigma^2$ obtained by both benchmarks are decreasing as the increase of $N$. This is due to the fact that more deployment of antennas provides more degree of freedom to align the signals from the selected devices. Obtaining the aggregate beamforming vector via the optimization methods can enjoy the advantage of multiple antennas. In contrast, sample a vector from the complex unit sphere as the aggregate beamforming vector cannot exploit the merit of multiple antennas. Hence, the obtained MSE performance under the proposed method remains unchanged w.r.t. the number of antennas $N$. When $K$ grows, the obtained $\text{MSE}/\sigma^2$ by the proposed method and 'Benchmark1' is getting smaller, while it is unaffected under 'Benchmark2'. These phenomenons suggest that our proposed method and 'Benchmark1' can reap the benefit from the device diversity, while 'Benchmark2' cannot. By comparing the proposed method and 'Benchmark1', we can observe that our proposed method is inferior to 'Benchmark1', since the optimized beamforming vector enjoy the benefit of multiple antennas at each iteration is also a vector in the complex unit sphere. Besides, the gap is shrinking as the increase of $K$, which means that the proposed method can approach 'Benchmark1'. It is worth noting that the computation for two benchmarks to obtaining $\mathbf{m}$ is intensive, which is negligible in our proposed random beamforming design.

To verify the MSE performance improvement by the refined method, we compare the refined



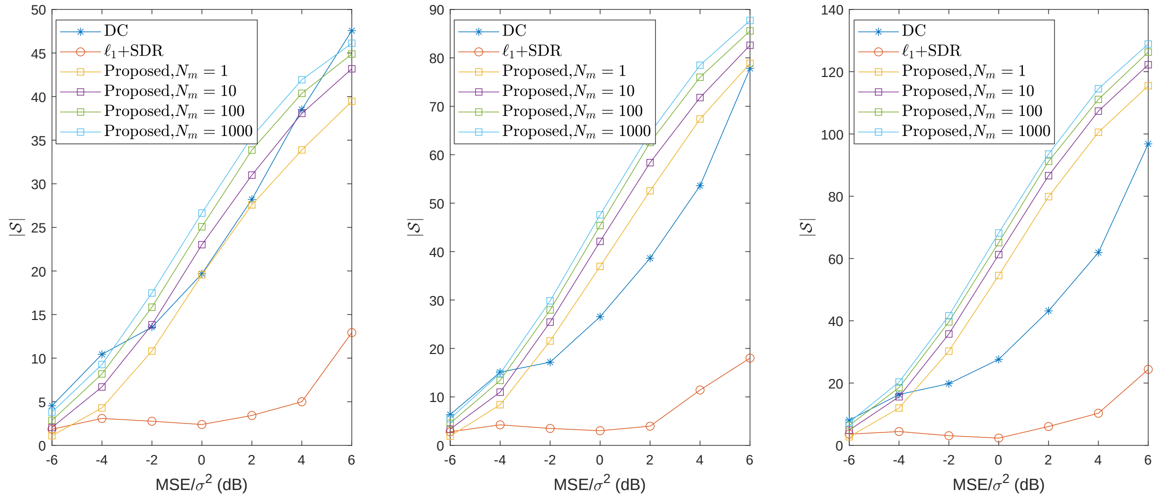
(a) (b)

Figure 8: MSE performance versus number of antennas. (a) Comparison of the proposed scheme with two benchmarks with different number of devices. (b) Proposed scheme with a fixed number of devices $K = 1000$ and different randomlization numbers.

methods with ‘Benchmark1’ in the scenario consisting of $K = 10^3$ devices and an aggregator with different number of antennas. Fig. 8(b) gives the MSE performance of ‘Benchmark1’ and the refined method under $N_m = 1, 10, 10^2, 10^3, 10^4$ randomizations. The figure shows that the MSE performance can be improved by sampling the vectors from the unit sphere for multiple times, which verifies the effectiveness of the proposed refined method. Besides, the performance improving rates become slower w.r.t. the number of randomizations $N_m$. When there are $N = 2, 4, 6$ antennas at the aggregator, the performance under $N_m = 10$, $10^2$ and $10^3$ randomizations can respectively achieve better performance than ‘Benchmark1’. A larger $N_m$ is required in order to get the performance outperforming ‘Benchmark1’ if more antennas are deployed. This can be explained such that the space of the unit sphere is much larger as the increase of $N$.

### C. Maximum Number of Selected Devices Performance

This subsection presents the performance of the number of the selected devices w.r.t. the threshold $\overline{\text{MSE}}/\sigma^2$, the range of which is set $[-6, +6]$ dB. We consider three settings where there are $N = 4$ antennas at the aggregator and $K = 50, 100,$ and $150$ devices in the system. We compare the proposed random aggregate beamforming based and the refined methods with the the novel DC and $\ell_1$+SDR algorithms, which are label as ‘DC’ and ‘$\ell_1$+SDR’ respectively. As shown in Fig. 9, more devices are selected if generating more random aggregate beamforming vectors,



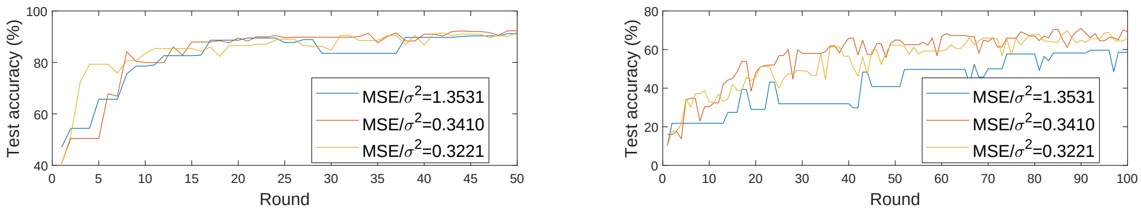
(a) (b) (c)

Figure 9: Number of selected devices $|\mathcal{S}|$ versus MSE with different number of devices. (a) $K = 50$. (b) $K = 100$. (c) $K = 150$.

which is referred to as the refined method. However, the improvement rate becomes slower as the increase of $N_m$. The $\ell_1$+SDR method is shown to have the poorest performance except for the case of $\overline{\text{MSE}}/\sigma^2 = -6$ dB. When the threshold is set to $-6$ dB, the $\ell_1$+SDR method achieves slightly better performance than our proposed random aggregate beamforming based method, but shows inferior performance compared with the proposed refined one. The number of selected devices $|\mathcal{S}|$ shows a growing trend w.r.t. the threshold $\overline{\text{MSE}}/\sigma^2$. In a small $\overline{\text{MSE}}/\sigma^2$ region, 'DC' method achieves the best performance compared with the other two methods. In the intermediate values of $[-2, +2]$, the proposed method outperforms 'DC' method, the gap of which is shrunk in the large region. Additionally, Fig. 9 shows that the advantage of our proposed method becomes more evident when there are more devices. It is worth noting here that the computational complexity of our proposed method is $\mathcal{O}(N_m K)$, which is negligible compared with both 'DC' and '$\ell_1$+SDR' reference methods.

### D. Learning Performance

The previous subsection presents the performance from the communication perspective. This subsection illustrates the impact of the MSE and the selected devices on the learning performance. We utilize FL framework to train the standard image classification tasks on two well-known



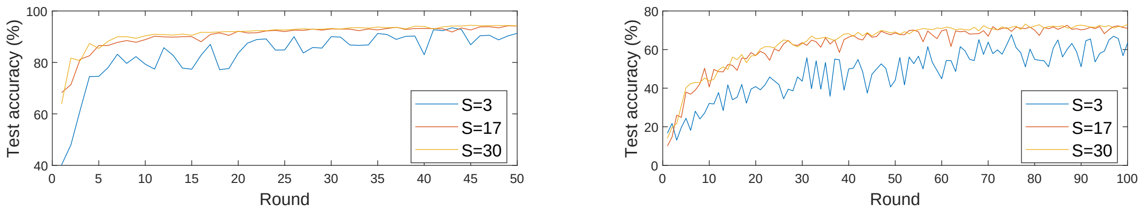
(a) (b)

Figure 10: Impact of MSE performance on the test accuracy averaged. (a) Non-i.i.d. MNIST. (b) Non-i.i.d. CIFAR10.

datasets, i.e., MNIST10 and CIFAR10. MNIST10 dataset consisting of $10$ classes of black-and-white handwritten digital picture is easier to learn, where a multilayer perception (MLP) neural network is adopted. By contrast, the color pictures of CIFAR10 are much difficult to learn, and thereby a more complex ResNet18 neural network is used. For simplicity, the effect of the aggregate error on the learning performance is modelled as the model retransmission, the probability of which is $p = 1 - \exp(-a\mathrm{MSE}/\sigma^2)$ where parameter $a$ is set to $1$.

For the impact of MSE on the learning performance, we consider the scenario where $S = 10$ devices are selected from $K = 100$ devices for model aggregation at the aggregator equipped with $N = 4$ antennas. The obtained average MSE performance $\mathrm{MSE}/\sigma^2$ under the proposed method, ‘Benchmark1’ and ‘Benchmark2’ methods are respectively $0.3221$, $0.3410$ and $1.3531$. Fig. 10 presents the testing accuracy on both datasets, which are non-i.i.d. distributed among the devices, which shows the slowest convergence rate under ‘Benchmark2’. This can be explained such that a smaller $\mathrm{MSE}/\sigma^2$ leads to a smaller retransmission probability and more validate training rounds. Since the MSE performance under our proposed method and ‘Benchmark1’ have approximately the same value, they have almost the same validate training epochs and thereby approximately the same test accuracy.

We also conduct the simulations on FL performance in terms of the number of selected devices under different methods. There are $K = 100$ devices, $N = 4$ antennas at the aggregator. The MSE performance threshold is set to $\overline{\mathrm{MSE}}/\sigma^2 = -2$ dB. In this case, the number of selected devices under ‘DC’, ‘$\ell_1$+SDR’ and the refined methods with $N_m = 10^3$ are $17$, $3$ and $30$ respectively. Fig. 11 displays the testing accuracy on non-i.i.d. MNIST and CIFAR10 datasets. As can be seen, the classification accuracy converges faster if more devices are selected for local model update and global model aggregation. This can be explained such that the average gradient computed



(a)

(b)

Figure 11: Impact of $|\mathcal{S}|$ on the test accuracy. (a) Non-i.i.d. MNIST. (b) Non-i.i.d. CIFAR10.

by a small number of devices are may far from the true one, which increases the performance fluctuation and slows the convergence rates.

## VII. Conclusion

In this paper, we investigate edge intelligence in a large-scale wireless network, where the FL framework is adopted to train a shared model. To aggregate the local models, the AirComp technique is adopted, which aggregates the local models in an analog manner. The design of device selection and model transmission schemes is critical for both learning and communication performances. We studied joint device selection and aggregate beamforming design with the two objectives of the aggregate error minimization and the number of selected devices maximization. To ease the computational complexity in a large-scale system, we proposed a random aggregate beamforming-based scheme, the core idea is to sample a vector from a complex unit sphere first and select the devices afterwards. When the number of the devices goes to infinity, we theoretically proved that the performance obtained by the proposed methods approached the optimal MSE performance for the aggregate error minimization problem. For the number of the selected devices maximization, the analysis also gave the interval and the average value of the number of the selected devices. When the number of devices is much smaller than infinity, the refined method was proposed aiming at performance improvement, the effectiveness of which was analyzed. Simulation results demonstrated the effectiveness of the proposed methods, and confirmed the theoretical analysis.

## References

[1] 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, Jan. 2020.

[2] Y. Huang, S. Liu, C. Zhang, X. You, and H. Wu, “True-data testbed for 5G/B5G intelligent network,” *Intell. Converged Networks*, vol. 2, no. 2, pp. 133–149, Jun. 2021.

[3] C. Yin, J. Xi, R. Sun, and J. Wang, “Location privacy protection based on differential privacy strategy for big data in industrial internet of things,” *IEEE Trans. Ind. Inf.*, vol. 14, no. 8, pp. 3628–3636, Aug. 2018.

[4] L. Guo, C. Zhang, J. Sun, and Y. Fang, “A privacy-preserving attribute-based authentication system for mobile health networks,” *IEEE Trans. Mob. Comput.*, vol. 13, no. 9, pp. 1927–1941, Sept. 2014.

[5] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[7] M. Chen, Z. Yang, W. Saad, C. Yin, S. Cui, and H. V. Poor, “A joint learning and communications framework for federated learning over wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[8] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, “Energy efficient federated learning over wireless communication networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2019.

[9] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” 2018. [Online]. Available: https://arxiv.org/abs/1706.02677

[10] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, “Reliable federated learning for mobile networks,” *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.

[11] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, “Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory,” *IEEE Int. Things J.*, vol. 6, no. 6, pp. 10 700–10 714, Dec. 2019.

[12] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, “Towards federated learning at scale: System design,” 2019. [Online]. Available: https://arxiv.org/abs/1902.01046

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” 2020. [Online]. Available: https://arxiv.org/abs/1812.06127

[14] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, “Scheduling policies for federated learning in wireless networks,” *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.

[15] T. Nishio and R. Yonetani, “Client selection for federated learning with heterogeneous resources in mobile edge,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–7.

[16] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, “TernGrad: Ternary gradients to reduce communication in distributed deep learning,” in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1508–1518.

[17] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, “Deep gradient compression: Reducing the communication bandwidth for distributed training,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada, May 2018.

[18] S. U. Stich, “Local SGD converges fast and communicates little,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–17.

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

[20] W. Shi, S. Zhou, and Z. Niu, “Device scheduling with fast convergence for wireless federated learning,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–6.

[21] G. Zhu, J. Xu, K. Huang, and S. Cui, “Over-the-air computing for wireless data aggregation in massive IoT,” 2020. [Online]. Available: https://arxiv.org/abs/2009.02181v1

[22] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, “Over-the-air computation for IoT networks: Computing multiple functions with antenna arrays,” *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5296–5306, Dec. 2018.

[23] M. Mohammadi 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, Mar. 2020.

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

[25] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, “Learning rate optimization for federated learning exploiting over-the-air computation,” 2021. [Online]. Available: https://arxiv.org/abs/2102.02946

[26] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1508–1518.

[27] N. Zhang and M. Tao, “Gradient statistics aware power control for over-the-air federated learning,” *IEEE Trans. Wireless Commun.*, in press, (DOI: 10.1109/TWC.2021.3065748).

[28] Z. He, A. S. Rakin, and D. Fan, “Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack,” in *Proc. Int. Conf. Comput. Vision Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 588–597.

[29] L. Xiao, Z. Zhang, and Y. Peng, “Noise optimization for artificial neural networks,” 2021. [Online]. Available: https://arxiv.org/abs/2102.04450

[30] W. Wen, Y. Wang, F. Yan, C. Xu, C. Wu, Y. Chen, and H. Li, “Smoothout: Smoothing out sharp minima to improve generalization in deep learning,” 2018. [Online]. Available: https://arxiv.org/abs/1805.07898

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[32] L. Chen, X. Qin, and G. Wei, “A uniform-forcing transceiver design for over-the-air function computation,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 942–945, Dec. 2018.

[33] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A first course in order statistics*. Philadelphia, PA, USA: SIAM, 2008.

[34] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Process Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.