SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2022, Vol. 65 132306:1–132306:9
https://doi.org/10.1007/s11432-020-3251-9

Robust federated learning for edge-intelligent networks

Zhihe GAO, Xiaoming CHEN* & Xiaodan SHAO

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027*, China*

Received 23 December 2020/Revised 17 March 2021/Accepted 21 April 2021/Published online 14 February 2022

Abstract The rapid development of machine learning and wireless communication is creating a new paradigm for future networks, namely edge-intelligent networks. Specifically, data generated by terminal devices is processed via machine learning at the edge of wireless networks, but not at the cloud. Owing to the growing concern for privacy information sharing, federated learning, as a new branch of machine learning, is appealing in edge-intelligent networks. For federated learning, the wireless transmission capabilities under practical conditions, e.g., imperfect channel state information (CSI), have a great impact on the accuracy of global aggregation of local model updates. Therefore, it is very important to enhance the robustness of communication for federated learning. In order to realize robust communication in the presence of channel uncertainty, we propose a robust federated learning algorithm for edge-intelligent networks, including device selection, transmit power allocation, and receive beamforming. Simulation results validate the robustness and effectiveness of the proposed robust federated learning algorithm in edge-intelligent networks.

Keywords edge-intelligent network, federated learning, imperfect CSI, robust design

Citation Gao Z H, Chen X M, Shao X D. Robust federated learning for edge-intelligent networks. Sci China Inf Sci, 2022, 65(3): 132306, https://doi.org/10.1007/s11432-020-3251-9

1 Introduction

In order to provide various advanced wireless services, the wireless network is undergoing a paradigm shift from traditional cloud computing architectures to mobile edge computing architectures [1, 2]. Through mobile edge computing, a large volume of data generated by massive terminal devices do not need to be sent to a data server at the cloud, but are processed at the edge of wireless networks, e.g., the base station (BS). Hence, the transmission latency can be decreased significantly. Recently, machine learning techniques are applied to mobile edge computing to improve the processing capability of wireless networks. As a result, wireless networks evolve to edge-intelligent networks [3, 4].

To realize edge intelligence based on machine learning, the BS needs to build a computation model. In general, there are two approaches to build the computation model at the BS, namely data sharing and model sharing. Specifically, data sharing collects the data from terminal devices to train the model based on deep learning, while model sharing exchanges model parameters between the BS and the terminal devices based on federated learning [5]. Since model sharing can protect the data privacy and reduce the communication burden, it is commonly adopted in edge-intelligent networks. Generally speaking, model sharing based on federated learning iteratively updates the model at the BS by averaging the parameters trained at the terminal devices. The authors in [6] proposed a practical method called federated averaging (FedAvg) for the federated learning based on iterative model averaging by using the aggregation of locally updated parameters with non-independent and identically distribution. Ref. [7] proposed some practical scheduling policies to further improve the performance of federated learning in wireless networks. Ref. [8] computed the summation part of the target function utilizing the superposition property of wireless channels. To accelerate the model aggregation, Ref. [9] proposed a novel federated learning framework based on over-the-air computation (AirComp). By exploiting the signal superposition property of a

* Corresponding author (email: chen_xiaoming@zju.edu.cn)

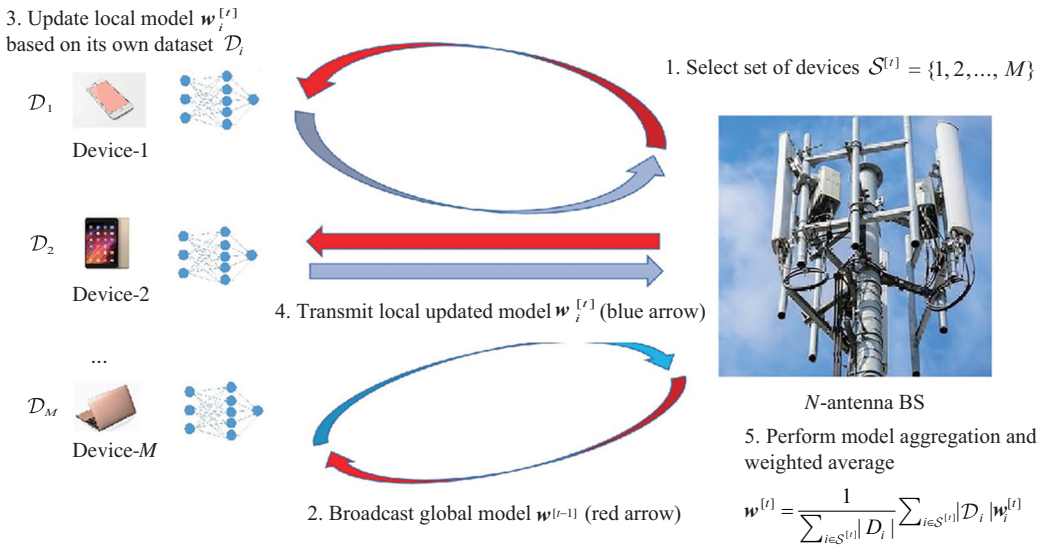

Figure 1 (Color online) Model update process of federated averaging (FedAvg) algorithm in t-round.

wireless multiple-access channel, AirComp can realize accurate aggregation via jointly designing transmit and receive schemes [10–13]. Especially, since the BS is usually equipped with multiple antennas, it is possible to further improve the aggregation accuracy. For AirComp-based federated learning over fading channels, the collection of participated devices has a great impact on the aggregation accuracy and the running efficiency. Ref. [14] proposed a device selection algorithm based on a metric termed as the age of update. Furthermore, the authors in [15] investigated the impact of noise in communications on the performance of federated learning. Previous studies commonly assumed full channel state information (CSI), to design AirComp-based federated learning algorithms. In practice, it is difficult to obtain full CSI about a large number of terminal devices and the aggregation error will become larger due to channel uncertainty. It has been proved in [16] that the aggregation error caused a notable drop of the prediction accuracy of federated learning. Hence, it is necessary to design robust federated learning algorithms to minimize the aggregation error in the presence of channel uncertainty.

In this paper, we consider a practical edge-intelligent network where the BS only has partial CSI by estimation or feedback. In such an adverse but practical scenario, we investigate the federated learning algorithm. The contributions of this paper are as follows:

(1) We provide an AirComp-based federated learning framework for edge-intelligent networks in the presence of imperfect CSI.

(2) We propose a robust federated learning algorithm, including device selection, transmit power allocation and receive beamforming to achieve accurate model aggregation.

The rest of this paper is organized as follows: Section 2 gives a brief introduction of an edge-intelligent network. A robust federated learning algorithm based on AirComp is proposed in Section 3. Simulation results are provided in Section 4. Finally, Section 5 concludes the paper.

Notations. Let upper (lower) case letters denote matrices (column vectors), $(\cdot)^{\mathrm{H}}$ denote conjugate transpose, $\|\cdot\|$ denote the L_2-norm of a vector, $|\cdot|$ denote the absolute value, $\mathbb{E}\{\cdot\}$ denote the expectation value.

2 System model

We consider an edge-intelligent network, which comprises a BS equipped with N antennas and K single-antenna intelligent devices, as shown in Figure 1. The BS deploys an edge server, which can iteratively construct a computation model based on the local models trained at the intelligent devices. During each iteration, the BS adopts the Fedavg algorithm [6] to update the computation model as follows.

At the t-th round iteration, the BS first selects a part of devices $\mathcal{S}^{[t]}$ from all devices with a given condition; then, the BS sends the global models $\boldsymbol{w}^{[t-1]}$ currently being trained to the selected devices; next, the i-th selected device updates its local model based on its own dataset $\mathcal{D}_i$ with the following

stochastic gradient descent method:

$$\boldsymbol{w}_i^{[t]} = \boldsymbol{w}^{[t-1]} - \theta_i \bigtriangledown L_i(\boldsymbol{w}^{[t-1]}, \mathcal{D}_i),$$

where $\bigtriangledown$ implies the gradient operator, θ_i is the learning rate, and $L_i(\cdot)$ denotes the loss function. Then, the device transmits $\boldsymbol{w}_i^{[t]}$ to the BS; finally, the BS performs model aggregation and weighted average according to their size of the dataset $|\mathcal{D}_i|$ to generate an updated global model $\boldsymbol{w}^{[t]}$. Thus, the updated aggregation model during the t-th iteration is given by

$$\boldsymbol{w}^{[t]} = \frac{1}{\sum_{i\in\mathcal{S}^{[t]}}|\mathcal{D}_i|} \sum_{i\in\mathcal{S}^{[t]}} |\mathcal{D}_i| \boldsymbol{w}_i^{[t]}, \tag{1}$$

where $\boldsymbol{w}_i^{[t]}$ of dimension d is the i-th device's local model at the current round, $|\mathcal{D}_i|$ is the pre-processing function at the i-th device, and $\frac{1}{\sum_{i\in\mathcal{S}^{[t]}}|\mathcal{D}_i|}$ is the post-processing function at the BS. In fact, the model aggregation in (1) can be realized by AirComp. Without loss of generality, we only study the situation in one round and omit the iteration index t. Specifically, let $\boldsymbol{s}_i = |\mathcal{D}_i|\boldsymbol{w}_i \in \mathbb{C}^d$, which is the i-th device's transmit signal with unit variance, i.e., $\mathbb{E}(\boldsymbol{s}_i\boldsymbol{s}_i^{\mathrm{H}}) = \mathbf{I}$ for ease of analysis. Each item in the vector $\boldsymbol{s}_i$ is sent to the BS sequentially over time slots. For simplification, we focus on one item transmission of the vector $\boldsymbol{s}_i$ and write $\boldsymbol{s}_i$ as s_i, which is independent and identically distributed (i.i.d.). According to the principle of AirComp, the aggregation signal at the BS can be expressed as

$$\boldsymbol{y} = \sum_{i\in\mathcal{S}} \boldsymbol{h}_i\sqrt{p_i}s_i + \boldsymbol{n}, \tag{2}$$

where p_i is the transmit power of the i-th device, $\boldsymbol{h}_i \in \mathbb{C}^N$ is the channel vector from the i-th device to the BS, and $\boldsymbol{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is the noise vector with σ^2 being the noise variance. With the aggregation signal, the BS utilizes a receive beamforming vector $\boldsymbol{z} \in \mathbb{C}^N$ to recover the desired aggregation model. Therefore, the actual aggregation model can be expressed as

$$\hat{m} = \boldsymbol{z}^{\mathrm{H}}\boldsymbol{y} = \boldsymbol{z}^{\mathrm{H}} \sum_{i\in\mathcal{S}} \boldsymbol{h}_i\sqrt{p_i}s_i + \boldsymbol{z}^{\mathrm{H}}\boldsymbol{n}. \tag{3}$$

According to (1), $m = \sum_{i\in\mathcal{S}} s_i$ is the desired aggregation model. In general, one can evaluate the performance by the mean square error (MSE) between the desired aggregation and the real aggregation [17]. The MSE is given by

$$\begin{aligned}\mathrm{MSE}(m,\hat{m}) &= \mathbb{E}\{(\hat{m}-m)(\hat{m}-m)^{\mathrm{H}}\} \\ &= \sum_{i\in\mathcal{S}} (\boldsymbol{z}^{\mathrm{H}}\boldsymbol{h}_i\sqrt{p_i} - 1)(\boldsymbol{z}^{\mathrm{H}}\boldsymbol{h}_i\sqrt{p_i} - 1)^{\mathrm{H}} + \sigma^2\|\boldsymbol{z}\|^2 \\ &= \sum_{i\in\mathcal{S}} |\boldsymbol{z}^{\mathrm{H}}\boldsymbol{h}_i\sqrt{p_i} - 1|^2 + \sigma^2\|\boldsymbol{z}\|^2.\end{aligned}$$

From (4), it is known that the set of the selected devices $\mathcal{S}$, the transmit power p_i, and the receive beamforming $\boldsymbol{z}$ determine the accuracy of the aggregation model for federated learning. Hence, it makes sense to jointly perform device selection, transmit power allocation and receive beamforming according to CSI. However, the BS is difficult to obtain full CSI about a large number of devices. In practice, the BS only has partial CSI by estimation or feedback. In general, the real CSI $\boldsymbol{h}_i$ related to the i-th device can be modeled as [18]

$$\mathcal{H}_i \triangleq \{\boldsymbol{h}_i = \hat{\boldsymbol{h}}_i + \boldsymbol{e}_i | \|\boldsymbol{e}_i\| \leqslant \varepsilon_i\},$$

where $\hat{\boldsymbol{h}}_i$ is obtained CSI and $\boldsymbol{e}_i$ is the channel error vector, whose norm is bounded by a given radius ε_i, i.e., $\|\boldsymbol{e}_i\| \leqslant \varepsilon_i$. Due to channel uncertainty at the BS, it is desired to design a robust federated learning algorithm for edge-intelligent networks to guarantee the accuracy of the aggregation model in the worse case.

3 Robust design for federated learning

In this section, we design a robust federated learning algorithm by jointly optimizing the set of selected devices, the transmit power, and the receive beamforming in the presence of channel uncertainty. First, we rewrite the MSE expression in (4) as follows:

$$\mathrm{MSE}(\boldsymbol{p},\boldsymbol{z})=\sum_{i\in\mathcal{S}}|\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\sqrt{p_i}-1|^2+\sigma^2\|\boldsymbol{z}\|^2. \tag{4}$$

In order to exploit multiuser diversity gain for federated learning, we formulate the design problem as the maximization of the number of selected devices while the MSE is smaller than a given value. Mathematically, it can be described as the following optimization problem:

$$\max_{\mathcal{S},p_i,\boldsymbol{z}\in\mathbb{C}^N}|\mathcal{S}|\quad\text{s.t. }\mathrm{MSE}(\boldsymbol{p},\boldsymbol{z})\leqslant\eta,\ p_i\leqslant P_{\max,i}, \tag{5}$$

where η is the tolerable maximum aggregation error, and $P_{\max,i}$ is the maximum transmit power of the i-th device. Since the objective function is discrete, the optimization problem is non-convex, which is difficult to obtain the optimal solutions directly.

It is seen in (5) that the MSE expression consists of two parts, one is the sum of the influence of each device on the overall MSE, the other is the influence of noise. Our task is to make the first part have the largest number of subitems, while the MSE error does not exceed the given value. That is to say, for any device i, where $i=1,2,\ldots,K$, if $|\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\sqrt{p_i}-1|^2$ is small, it is better to select this device. However, due to the undetermined beamforming vector $\boldsymbol{z}$ and transmit power p_i, it is impossible to obtain the value of each $|\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\sqrt{p_i}-1|^2$. To solve this challenge, we first assume all devices participate in model aggregation for federated learning. Then, we derive the transmit power and receive beamforming by minimizing the maximum MSE in the presence of channel uncertainty. Finally, we select devices with small $|\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\sqrt{p_i}-1|^2$ until the total MSE is close to but not more than the given value η. In what follows, we design the robust federated learning algorithm according to the above idea.

3.1 Algorithm design

When all the K devices participate in model aggregation, the MSE can be represented as $\sum_{i=1}^{K}|\boldsymbol{z}^{\mathrm{H}}\boldsymbol{h}_i\sqrt{p_i}-1|^2+\sigma^2\|\boldsymbol{z}\|^2$. Due to channel uncertainty, we cannot obtain the minimum MSE, but the maximum MSE in the worse case. This leads to the following min-max problem:

$$\min_{\boldsymbol{z},p_i,\forall i}\ \max_{\boldsymbol{h}_i\in\mathcal{H}_i}\ \mathrm{MSE}(\boldsymbol{p},\boldsymbol{z}) \tag{7a}$$

$$\text{s.t. }p_i\leqslant P_{\max,i},\ \forall\boldsymbol{e}_i:\|\boldsymbol{e}_i\|^2\leqslant\varepsilon_i. \tag{7b}$$

Owing to the coupling of the two optimization variables in the MSE, the optimization problem (7) is not convex. To solve this problem, we adopt the alternating optimization (AO) approach [19]. Specifically, we iteratively optimize one variable by fixing the other variable until convergence. First, given the transmit power, we optimize the receive beamforming vector $\{\boldsymbol{z}\}$. To facilitate problem solving, we introduce an auxiliary variable α_i. The problem is transformed as

$$\min\ \sum_{i=1}^{K}\alpha_i+\sigma^2\|\boldsymbol{z}\|^2 \tag{8a}$$

$$\text{s.t. }|\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\sqrt{p_i}-1|^2\leqslant\alpha_i, \tag{8b}$$

$$\alpha_i\geqslant 0, \tag{8c}$$

$$\forall\boldsymbol{e}_i:\|\boldsymbol{e}_i\|^2\leqslant\varepsilon_i.$$

The constraint (8b) is non-convex due to channel uncertainty. In order to ensure the robustness and feasibility of the algorithm, we introduce the following lemmas to transform the constraint (8b).

Lemma 1 (Schur's complement [20])**.** Let $\boldsymbol{M}$ be a Hermitian matrix given by $\boldsymbol{M}=\left[\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B}^{\mathrm{H}}\\ \boldsymbol{B} & \boldsymbol{C}\end{smallmatrix}\right]$. Then, $\boldsymbol{M}$ is semi-positive, i.e., $\boldsymbol{M}\geqslant 0$, if and only if $\boldsymbol{A}-\boldsymbol{B}^{\mathrm{H}}\boldsymbol{C}^{-1}\boldsymbol{B}\geqslant 0$ with assuming $\boldsymbol{C}$ is invertible, or $\boldsymbol{C}-\boldsymbol{B}^{\mathrm{H}}\boldsymbol{A}^{-1}\boldsymbol{B}\geqslant 0$ with assuming $\boldsymbol{A}$ is invertible.

Lemma 2 ([21]). Let us define a matrix function $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{A} - (\boldsymbol{B}^{\mathrm{H}}\boldsymbol{x}\boldsymbol{c}^{\mathrm{H}} + \boldsymbol{B}\boldsymbol{x}^{\mathrm{H}}\boldsymbol{c})$, where $\boldsymbol{B} \in \mathbb{C}^{m\times n}$, $\boldsymbol{x} \in \mathbb{C}^{m\times 1}$, $\boldsymbol{c} \in \mathbb{C}^{n\times 1}$, and $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is a Hermitian matrix. Then

$$\boldsymbol{F}(\boldsymbol{x}) \geqslant 0,\ \forall \boldsymbol{x} : \|\boldsymbol{x}\| \leqslant \varepsilon$$

holds true if and only if there exits $\lambda \geqslant 0$, such that

$$\begin{bmatrix} \boldsymbol{A} - \lambda \boldsymbol{c}\boldsymbol{c}^{\mathrm{H}} & -\varepsilon \boldsymbol{B}^{\mathrm{H}} \\ -\varepsilon \boldsymbol{B} & \lambda \mathbf{I} \end{bmatrix} \geqslant 0.$$

According to Lemma 1, the constraint (8b) can be rewritten as

$$\begin{bmatrix} \alpha_i & (\sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1)^{\mathrm{H}} + \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\boldsymbol{e}_i \\ \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1 + \sqrt{p_i}\boldsymbol{e}_i^{\mathrm{H}}\boldsymbol{z} & 1 \end{bmatrix} \geqslant 0.$$

Let

$$\boldsymbol{A}_i = \begin{bmatrix} \alpha_i & (\sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1)^{\mathrm{H}} \\ \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1 & 1 \end{bmatrix}, \quad \boldsymbol{B}_i = \begin{bmatrix} \mathbf{0} & -\sqrt{p_i}\boldsymbol{z} \end{bmatrix},$$

and $\boldsymbol{c} = [1\ \mathbf{0}]^{\mathrm{T}}$, then

$$\begin{aligned} &\begin{bmatrix} \alpha_i & (\sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1)^{\mathrm{H}} + \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\boldsymbol{e}_i \\ \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1 + \sqrt{p_i}\boldsymbol{e}_i^{\mathrm{H}}\boldsymbol{z} & 1 \end{bmatrix} \\ &\quad = \boldsymbol{A}_i - (\boldsymbol{B}_i^{\mathrm{H}}\boldsymbol{e}_i\boldsymbol{c}^{\mathrm{H}} + \boldsymbol{c}_i\boldsymbol{e}_i^{\mathrm{H}}\boldsymbol{B}_i) \geqslant 0, \quad \forall \boldsymbol{e}_i : \|\boldsymbol{e}_i\| \leqslant \varepsilon_i. \end{aligned}$$

Further, according to Lemma 2, we have

$$\begin{aligned} &\boldsymbol{A}_i - (\boldsymbol{B}_i^{\mathrm{H}}\boldsymbol{e}_i\boldsymbol{c}^{\mathrm{H}} + \boldsymbol{c}\boldsymbol{e}_i^{\mathrm{H}}\boldsymbol{B}_i) \geqslant 0, \quad \forall \boldsymbol{e}_i : \|\boldsymbol{e}_i\| \leqslant \varepsilon_i \\ &\iff \begin{bmatrix} \boldsymbol{A}_i - \phi_i \boldsymbol{c}\boldsymbol{c}^{\mathrm{H}} & -\varepsilon_i \boldsymbol{B}_i^{\mathrm{H}} \\ -\varepsilon_i \boldsymbol{B}_i & \phi_i \mathbf{I} \end{bmatrix} \geqslant \mathbf{0}, \quad \exists \phi_i \geqslant 0. \end{aligned} \tag{10}$$

Thus, the constraint (8b) is converted to

$$\begin{bmatrix} \alpha_i - \phi_i & (\sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1)^{\mathrm{H}} & \mathbf{0} \\ \sqrt{p_i}\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{h}}_i - 1 & 1 & \varepsilon_i\sqrt{p_i}\boldsymbol{z}^{\mathrm{H}} \\ \mathbf{0} & \varepsilon_i\sqrt{p_i}\boldsymbol{z} & \phi_i \mathbf{I} \end{bmatrix} \geqslant \mathbf{0}, \quad \exists \phi_i \geqslant 0, \tag{11}$$

which is a linear matrix inequality (LMI) and thus is convex. In this case, the problem (8) can be transformed as

$$\begin{aligned} &\min_{\boldsymbol{z},\phi_i,\forall i} \ \sum_{i=1}^{K} \alpha_i + \sigma^2\|\boldsymbol{z}\|^2 \\ &\text{s.t.} \quad \alpha_i \geqslant 0,\ \phi_i \geqslant 0 \text{ and (11)}. \end{aligned} \tag{12}$$

Problem (12) is convex and thus can be solved by some optimization tools, e.g., CVX. Second, we assume that the receive beamforming vector is fixed to optimize the transmit power $\{p_i\}$. In this case, the problem (8) can be transformed as

$$\begin{aligned} &\min_{\alpha_i,p_i,\forall i} \ \sum_{i=1}^{K} \alpha_i + \sigma^2\|\boldsymbol{z}\|^2 \\ &\text{s.t.} \quad |\boldsymbol{z}^{\mathrm{H}}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\sqrt{p_i} - 1|^2 \leqslant \alpha_i,\ \alpha_i \geqslant 0,\ p_i \leqslant P_{\max,i}. \end{aligned} \tag{13}$$

In order to solve this problem, we also need to transform the constraint (8b). Let $P_i = \sqrt{p_i}$, the constraint (8b) can be transformed as

$$\begin{bmatrix} \alpha_i - \phi_i & P_i \boldsymbol{z}^{\mathrm{H}} \hat{\boldsymbol{h}}_i - 1 & \mathbf{0} \\ P_i \boldsymbol{z}^{\mathrm{H}} \hat{\boldsymbol{h}}_i - 1 & 1 & \varepsilon_i P_i \boldsymbol{z}^{\mathrm{H}} \\ \mathbf{0} & \varepsilon_i P_i \boldsymbol{z} & \phi_i \mathbf{I} \end{bmatrix} \succcurlyeq \mathbf{0}, \quad \exists \phi_i \geqslant 0. \tag{14}$$

Hence, the problem (13) is transformed as

$$\begin{aligned} &\min_{\alpha_i, p_i, \forall i} \sum_{i=1}^{K} \alpha_i + \sigma^2 \|\boldsymbol{z}\|^2 \\ &\text{s.t.} \quad \phi_i \geqslant 0,\ P_i^2 \leqslant\ p_i,\ \ p_i \leqslant P_{\max,i},\ \text{(8b)}, \text{(8c) and (14)}. \end{aligned} \tag{15}$$

Problem (15) is also convex and thus can be solved by some optimization tools directly. By iteratively optimizing the above two problems until convergence, we can obtain the transmit power p_i, receive beamforming $\boldsymbol{z}$, and the corresponding MSE $\alpha_i, \forall i$. Without loss of generality, we assume α_i has an ascending order, namely $\alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_K$. Then, we select the first k^* devices as the ones participating in federated learning as follows:

$$k^* = \arg\max_{k} \ \left\{ k \mid \sum_{i=1}^{k} \alpha_i \leqslant \eta \right\}.$$

In summary, the proposed robust federated learning algorithm can be described as Algorithm 1.

Algorithm 1 Robust federated learning for edge-intelligent networks

Input: Number of antennas N, number of total devices K, obtained CSI $\hat{\boldsymbol{h}}_i$, maximum transmit power $P_{\max,i}$, channel error vector's norm bound ε_i, noise power σ^2, tolerance of MSE η;
Output: The set of selected devices $\mathcal{S}$, beamforming vector $\boldsymbol{z}$, transmit power p_i, $i \in \mathcal{S}$;
1: Initialize $\boldsymbol{z}^{(0)}$, $p_i^{(0)} = P_{\max,i}/2$, and iteration index $t = 1$;
2: Repeat
3: Obtain $\boldsymbol{z}^{(t)}$ by solving problem (12) via CVX with fixed $p_i^{(t-1)}$;
4: Obtain $p_i^{(t)}$ by solving problem (15) via CVX with fixed $\boldsymbol{z}^{(t)}$;
5: $t = t + 1$;
6: Until convergence;
7: Obtain α_i, $i = 1, 2, \ldots, K$, sort them and obtain $\alpha_{\pi(i)}$, $i = 1, 2, \ldots, K$, $\alpha_{\pi(1)} \leqslant \alpha_{\pi(2)} \leqslant \cdots \leqslant \alpha_{\pi(K)}$;
8: Add $\pi(i)$-th device to $\mathcal{S}$, until $\sum_{i=1}^{L} \alpha_{\pi(i)} + \sigma^2 \|\boldsymbol{z}\|^2 \leqslant \eta$ and $\sum_{i=1}^{L+1} \alpha_{\pi(i)} + \sigma^2 \|\boldsymbol{z}\|^2 \geqslant \eta$;
9: Obtain $\mathcal{S}$=$\{\pi(1), \ldots, \pi(L)\}$.

3.2 Convergence and complexity analysis

In this subsection, we analyze the convergence behavior and computational complexity of the proposed algorithm.

For Algorithm 1, it converges as long as the initial value is set properly. First, since the problem (11) is convex in terms of $\{\boldsymbol{z}^{(t)}\}$, it is feasible to find the optimal solutions for minimizing the objective value via CVX directly. Then, since the problem (15) is convex in terms of $\{p_i^{(t)}\}$, it is also feasible to find the optimal solutions via CVX. Thus, based on the steps in Algorithm 1, the solutions in the t-th iteration are feasible for the original problem (12) in the $(t+1)$-th iteration, which means that the objective value obtained in the $(t+1)$-th iteration is less than that in the t-th iteration. In other words, the MSE monotonically decreases after each iteration. Furthermore, due to the existence of the transmit power constraints $p_i \leqslant P_{\max,i}$ at each device, the MSE is lower bounded. According to the monotone bounded convergence theorem, Algorithm 1 is convergent. The proposed algorithm adopts the AO approach. According to [22, 23], the convergence rate shows a two-stage behavior. At first, the objective function decreases q-linearly until sufficiently small. After that, sub-linear convergence is initiated.

Since the computational complexity of each iteration is the same, we only analyze the per-iteration complexity in the following. In each iteration, the computational complexity of solving problem (12) is dominant [21]. By using CVX to solve the problem (12), the CVX tool employs a standard interior-point

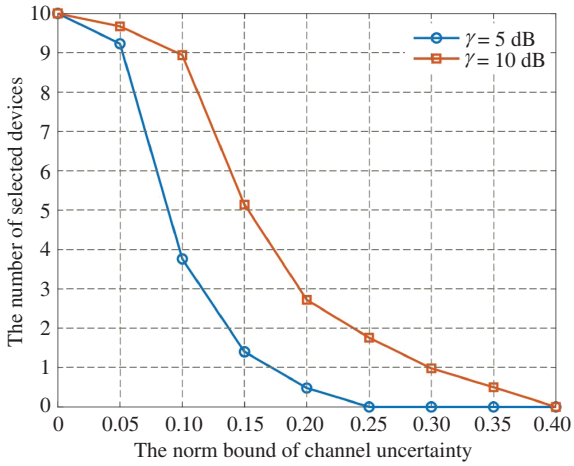

Figure 2 (Color online) The number of selected devices with different norm bounds of channel uncertainty ε. The number of total devices K is 10 and the number of BS antennas N is 32. The MSE requirement γ is 5 and 10 dB, respectively.

method (IPM) [24]. The complexity of this method depends on the constraints. Specifically, it has K LMI constraints of size 1, K LMI constraints of size $N+2$. Thus, for a given precision $\varepsilon > 0$ of solution, the per-iteration complexities of solving problem (12) by IPM is $\ln \frac{1}{\varepsilon} = \sqrt{K(N+3)} \cdot n \cdot [K\big(1+(N+2)^3\big) + Kn\big(1+(N+2)^2\big)]$, where the decision variable n is in the order of $\mathcal{O}(KN^2)$. Moreover, the complexity of sorting algorithm is $\mathcal{O}(K \log_2 K)$.

4 Simulation results

In this section, we present simulation results to validate the robustness and effectiveness of the proposed algorithm for the model aggregation of federated learning in edge-intelligent networks. We assume that each device has the same maximum transmit power $P_{\max,i} = P_{\max}$ and the same norm bound of channel uncertainty $\varepsilon_i = \varepsilon$. We use $\text{SNR} = 10\log_{10}(P_{\max}/\sigma^2)$ to denote the transmit SNR (in dB). In the simulations, let all SNR = 20 dB and MSE requirement (in dB) be $\gamma = 10\log_{10}(P_{\max}\eta/\sigma^2)$. In practice, the value of γ should be determined according to the adopted training model and the required prediction accuracy.

First, we investigate the impact of channel uncertainty on device selection. As shown in Figure 2, as the norm bound of channel error vector increases, the number of selected devices decreases accordingly. This is because a larger channel uncertainty leads to a larger MSE, and thus a smaller number of devices are selected for model aggregation to guarantee the aggregation accuracy.

Then, we check the influence of the number of total devices K and the number of BS antennas N on the number of selected devices. As shown in Figure 3, the proposed algorithm selects more devices as the number of BS antennas increases. This is because more antennas decrease the MSE. On the other hand, an increase in the number of total devices may lead to a decrease in the number of selected devices. Since we assume that all devices are involved in model aggregation, the designed beamforming vector and transmit power are sub-optimal if only a part of devices are involved in model aggregation. As a result, an increase of total devices may increase the MSE. However, if there are enough BS antennas, an increase of total devices can lead to an increase of selected devices under a suitable MSE requirement. If the number of total devices is large, we can increase the number of selected devices by adding BS antennas.

To show the performance of the proposed algorithm for federated learning tasks, we further train a convolutional neural network (CNN) on the Mixed National Institute of Standards and Technology database (MNIST) with a 32-antenna edge server and 10 mobile devices. The MNIST database of handwritten digits has a training set of 60000 examples and a test set of 10000 examples. This CNN has two 5×5 convolution layers (the first with 32 channels, the second with 64, each followed with 2×2 max pooling), a fully connected layer with 512 units and ReLu activation, and a final softmax output layer (1663370 total parameters). The data is shuffled, and each device has 600 examples. In [9], the authors formulated the problem of device selection under perfect channel conditions as a nonconvex optimization problem with a sparse objective function and a low-rank constraint. The authors employed three algorithms to

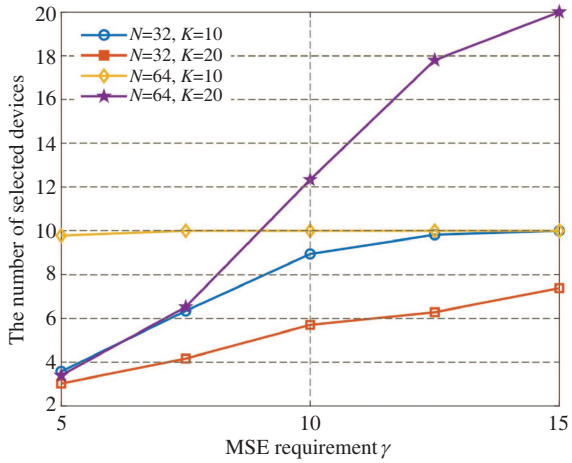

Figure 3 (Color online) The number of selected devices with different numbers of BS antennas and total devices. The norm bound of channel uncertainty ε is 0.1.

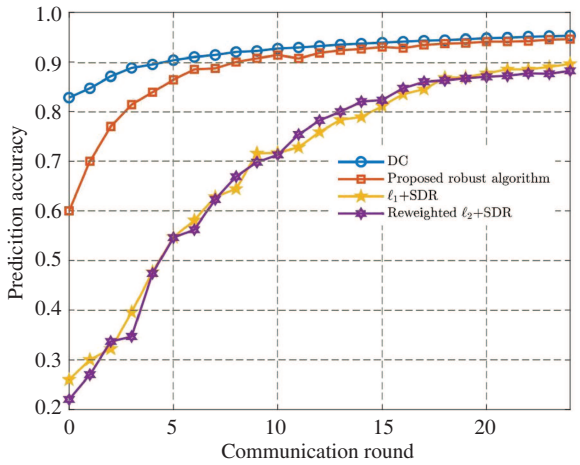

Figure 4 (Color online) Performance of different device selection algorithms in CNN training. The MSE requirement is 5 dB. The norm bound of channel uncertainty ε in the proposed robust algorithm is 0.1.

solve this problem and made a comparison. We compare the proposed robust algorithm with these three baseline algorithms, namely ℓ_1+semidefinite relaxation (SDR) [25], difference-of-convex-function (DC) [9] and reweighted ℓ_2+SDR [26]. Note that the baseline algorithms work with perfect CSI, but the proposed robust algorithm works in the presence of channel uncertainty. As shown in Figure 4, the proposed robust algorithm performs better than ℓ_1+SDR and reweighted ℓ_2+SDR. Even in the presence of imperfect CSI, the proposed algorithm achieves the same prediction accuracy as the DC algorithm even with a not so large number of communication rounds. In other words, the proposed algorithm has high robustness in federated learning. Hence, the proposed algorithm is appealing in practical edge-intelligent networks.

5 Conclusion

In this paper, we provided a framework of federated learning for edge-intelligent networks in the presence of channel uncertainty. A robust federated learning algorithm including device selection, power allocation and receive beamforming was proposed. Simulation results have shown its good performance in the training of CNN on the MNIST dataset. The proposed robust federated learning algorithm can be widely applied to widely practical edge-intelligent networks such as cellular internet of things [27, 28]. Thus, the processing latency and signaling overhead can be reduced significantly.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018YFB1801104), National Natural Science Foundation of China (Grant No. 61871344), and Zhejiang Provincial Natural Science Foundation of China (Grant No. LR20F010002).

References

1 Mao Y, You C, Zhang J, et al. A survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tut, 2017, 19: 2322–2358
2 Chen X, Qi Q. Convergence of Energy, Computation and Communication in B5G Cellular Internet of Things. Berlin: Springer, 2020
3 Chen X, Ng D W K, Yu W, et al. Massive access for 5G and beyond. IEEE J Sel Areas Commun, 2021, 39: 615–637
4 Wang K H, Xiong Z H, Chen L, et al. Joint time delay and energy optimization with intelligent overclocking in edge computing. Sci China Inf Sci, 2020, 63: 140313
5 Zhao Z, Feng C, Yang H H, et al. Federated-learning-enabled intelligent fog radio access networks: fundamental theory, key techniques, and future trends. IEEE Wireless Commun, 2020, 27: 22–28
6 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017. 54: 1273–1282
7 Yang H H, Liu Z, Quek T Q S, et al. Scheduling policies for federated learning in wireless networks. IEEE Trans Commun, 2020, 68: 317–333
8 Chen L, Zhao N, Chen Y, et al. Communicating or computing over the MAC: function-centric wireless networks. IEEE Trans Commun, 2019, 67: 6127–6138
9 Yang K, Jiang T, Shi Y, et al. Federated learning via over-the-air computation. IEEE Trans Wireless Commun, 2020, 19: 2022–2035
10 Amiri M M, Gunduz D. Machine learning at the wireless edge: distributed stochastic gradient descent over-the-air. IEEE Trans Signal Process, 2020, 68: 2155–2169
11 Amiri M M, Gunduz D. Federated learning over wireless fading channels. IEEE Trans Wireless Commun, 2020, 19: 3546–3557
12 Zhu G, Wang Y, Huang K. Broadband analog aggregation for low-latency federated edge learning. IEEE Trans Wireless Commun, 2020, 19: 491–506
13 Zhu G, Du Y, Gunduz D, et al. One-bit over-the-air aggregation for communication-efficient federated edge learning: design and convergence analysis. IEEE Trans Wireless Commun, 2021, 20: 2120–2135
14 Yang H H, Arafa A, Quek T Q S, et al. Age-based scheduling policy for federated learning in mobile edge networks. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, 2020. 1–5
15 Ang F, Chen L, Zhao N, et al. Robust federated learning with noisy communication. IEEE Trans Commun, 2020, 68: 3452–3464
16 Keskar N S, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: generalization gap and sharp minima. In: Proceedings of International Conference on Learning Representations (ICLR), 2017
17 Li C, Wang J, Zheng F C, et al. Overhearing-based co-operation for two-cell network with asymmetric uplink-downlink traffics. IEEE Trans Signal Inf Process over Networks, 2016, 2: 350–361
18 Wang J H, Palomar D P. Worst-case robust MIMO transmission with imperfect channel knowledge. IEEE Trans Signal Process, 2009, 57: 3086–3100
19 Qi Q, Chen X, Ng D W K. Robust beamforming for NOMA-based cellular massive IoT with SWIPT. IEEE Trans Signal Process, 2020, 68: 211–224
20 Boyd S, Vandenberghe L. Convex Optimization. Cambridge: Cambridge University Press, 2004
21 Qi Q, Chen X, Zhong C, et al. Integrated sensing, computation and communication in B5G cellular Internet of Things. IEEE Trans Wireless Commun, 2021, 20: 332–344
22 Bezdek J C, Hathaway R J. Convergence of alternating optimization. Neural Paral Sci Comput, 2003, 11: 351–368
23 Both J W. On the rate of convergence of alternating minimization for non-smooth non-strongly convex optimization in Banach spaces. 2019. ArXiv:1911.00404
24 Ben-Tal A, Nemirovski A. Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. Philadelphia: SIAM, 2001
25 Luo Z Q, Sidiropoulos N D, Tseng P, et al. Approximation bounds for quadratic optimization with homogeneous quadratic constraints. SIAM J Opt, 2007, 18: 1–28
26 Shi Y, Cheng J, Zhang J, et al. Smoothed l_p-minimization for green cloud-RAN with user admission control. IEEE J Sel Areas Commun, 2016, 34: 1022–1036
27 Chen X. Massive Access for Cellular Internet of Things Theory and Technique. Berlin: Springer, 2019
28 Qi Q, Chen X M, Zhong C J, et al. Physical layer security for massive access in cellular Internet of Things. Sci China Inf Sci, 2020, 63: 121301