Resource Consumption for Supporting Federated Learning Enabled Network Edge Intelligence

Yi-Jing Liu*, Gang Feng*, Yao Sun†, Xiaoqian Li*, Jianhong Zhou‡*, Shuang Qin*
*National Key Laboratory of Science and Technology on Communications,
and Yangtze Delta Region Institute (Huzhou),
University of Electronic Science and Technology of China
†James Watt School of Engineering, University of Glasgow
‡School of Computer and Software Engineering, Xihua University
E-mail:fenggang@uestc.edu.cn

***Abstract*—Federated learning (FL) has recently become one of the hottest focuses in network edge intelligence. In the FL framework, user equipments (UEs) train local machine learning (ML) models and transmit the trained models to an aggregator where a global model is formed and then sent back to UEs, such that FL can enable collaborative model training. In large-scale and dynamic edge networks, both local model training and transmission may not be always successful due to constrained power and computing resources at mobile devices, wireless channel impairments, bandwidth limitations, etc., which directly degrades FL performance in terms of model accuracy and/or training time. On the other hand, we need to quantify the benefits and cost of deploying edge intelligence when we plan to improve network performance by using artificial intelligence (AI) techniques which definitely incur certain cost. Therefore, it is imperative to deeply understand the relationship between the required multiple-dimensional resources and FL performance to facilitate FL enabled edge intelligence. In this paper, we construct an analytical model for investigating the relationship between the accuracy of ML model and consumed network resources in FL enabled edge networks. Based on the analytical model, we can explicitly quantify the trained model accuracy given spatial-temporal domain distribution, available user computing and communication resources. Numerical results validate the effectiveness of our theoretical modeling and analysis. Our analytical model in this paper provides some useful guidelines for appropriately promoting FL enabled edge network intelligence.**

I. Introduction

Edge intelligence is boosted by the unprecedented computing capability of smart devices. Nowadays, more than 10 billion Internet-of-Things (IoT) equipment and 5 billion smartphones have emerged that are equipped with artificial intelligence (AI)-empowered computing modules, such as AI chips and graphic processing units (GPUs) [1]. On the one hand, the user equipment (UE) can be potentially deployed as computing nodes to support emerging services, such as collaborative tasks, which paves the way for applying AI in wireless edge networks. On the other hand, in the paradigm of machine learning (ML), the powerful computing capability on these UEs can decouple conventional ML from acquiring, storing, and training data in data centers as conventional methods.

Federated learning (FL) has recently been widely acknowledged as one of the most essential enablers to bring edge intelligence into reality, as it facilitates collaborative training of ML models, while enhancing individual user privacy and data security [2], [3]. In FL, ML models are trained locally, therefore raw data remains in the device. Specifically, FL uses an iterative approach that requires a number of global iterations to achieve a certain global model accuracy. In each global iteration, UEs perform several local iterations to reach a local model accuracy [2], [3]. As a result, the implementation of FL in wireless networks can also reduce the costs of transmitting raw data, relieve the burden on backbone networks, and reduce latency for real-time decisions, etc.

While FL offers these attractive and valuable benefits, it also faces many challenges, especially when being deployed in wireless edge networks. For example, both local training and model transmission can be unsuccessful due to constrained resources and unstable transmission. Moreover, different from the conventional ML approaches, where raw datasets are sent to a central server, only the lightweight model parameters (*i.e.*, weights, gradients, etc.) are exchanged in FL. Nevertheless, the communication cost of FL could be still fairly large and cannot be ignored. The experimental results in [4] show that the model size of a 5-layer convolutional neural network used for MNIST (classification) is about 4.567MB per global iteration for 28×28 images. Therefore, before deploying FL empowered wireless edge networks, we need to answer two fundamental questions: (1) How accurate of an ML model can be achieved by using FL, and (2) How much cost is incurred to guarantee certain required FL performance? Obviously, answering these two questions is of paramount importance for facilitating edge network intelligence. Therefore, we need to deeply understand the relationship between FL performance and consumed multi-dimensional resources.

In this paper, we theoretically analyze how many resources

This work has been supported by the Key Research and Development Projects (Grant 2020YFB1806804), Huawei Cooperation Projects (Grant TC20210316002), and Basic Business Fees for Central Colleges and Universities (Grant ZYGX2020ZB044).

are needed to support an FL-enabled edge intelligent network. The main contributions of this paper can be summarized as follows. (1) We develop an analytical model for FL empowered wireless edge networks, where UE geographical distribution and arrival rate of the interfering UEs are modeled as Poisson Point Process (PPP). (2) We theoretically analyze SINR, SNR, and the local/global model transmission success probability. (3) We derive the explicit expression of the model accuracy under FL framework as a function of the amount of consumed resources. Based on this, we discuss three specific cases according to the sufficiency of respective communication and computing resources. Simulation results demonstrate the effectiveness of our theoretical modeling and analysis.

In the rest of this paper, we begin with the general FL enabled edge network model in Section II. Then we present analysis for consumed communication/computing resources in FL in Section III. The relationship between FL performance and consumed resources and different cases are presented in Section IV and Section V respectively. Finally, we present the numerical results in Section VI and conclude the paper in Section VII.

II. FL enabled Wireless Network Model

We consider an FL enabled radio access network (RAN) consisting of a central BS and multiple UEs, shown as Fig.1. The UEs can be regarded as local computing nodes while the server is the aggregator associated with the central BS [2], [5].

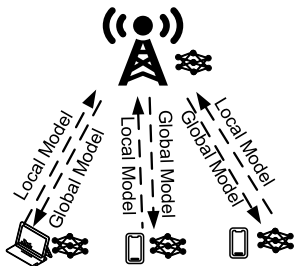

Fig. 1: The FL empowered wireless network.

A. FL Model

1) Loss Function: Assume that n is the number of UEs that are geographically distributed as homogeneous PPP with intensity λ_i. For a specific UE i, it has a local dataset $\mathcal{S}_i$ with S_i data samples, where $\mathcal{S}_i = \{x_{ik} \in \mathbb{R}^d, y_{ik} \in \mathbb{R}\}_{k=1}^{S_i}$. Moreover, we define $f_k(w_i; x_{ik}, y_{ik})$ as a loss function for data sample k of UE i to capture the FL performance which is different for various FL learning tasks [6]. For example, for a linear regress, the loss function is $f_k(w_i^r(t); x_{ik}, y_{ik}) = \frac{1}{2}(x_{ik}^{\mathrm{T}} w_i - y_{ik})^2$. Furthermore, we define $F_i(w) : \mathbb{R}^m \to \mathbb{R}$ as $F_i(w) \triangleq \frac{1}{S_i}\sum_{k\in S_i} f_k(w_i^r(t); x_{ik}, y_{ik})$, where $S \triangleq \sum_{i=1}^{n} S_i$. The goal of the BS is to fit a vector w so as to minimize $F(w)$, *i.e.*, $w^* \triangleq \arg_w \min F(w)$.

2) Updating Model: In FL, each global iteration is called a *communication round* [5]. A communication round consists of a number of phases including local model updating, local iterations, local model transmission, global model updating, and global model transmission. In the following, we present the details of the local and global model updating respectively.

(1) *Local Model Updating*: Based on the local learning algorithms (*e.g.*, gradient descent (GD), stochastic gradient descent (SGD), etc.), the local model $w_i^r(t)$ is updated as $w_i^r(t) = g_r - \eta \nabla F_i(w_i^r(t-1))$ when $0 < t \leq \tau$ and $w_i^r(t) = g_r$ when $t = 0$, where t is the number of local iterations, $\eta \geq 0$ is the step size and g_r is the global model at the r-th communication round.

(2) *Global Model Updating*: After τ local iterations, *i.e.*, $t = \tau$, UEs will achieve certain local accuracy and send the local model to the aggregator. Then the global aggregation is performed at the aggregator according to $g_r = \frac{1}{S}\sum_{i=1}^{n} S_i w_i^r(t), t = \tau$.

B. Computing Resource Consumption Model

For a specific UE i, let Z_i (cycles/s) and c_i (cycles/sample) denote its computing capacity and the number of CPU cycles required for computing one sample data at UE i respectively. T_i represents the local computing time needed for one local iteration. Therefore, the consumed computing resources during one local iteration for UE i is given by $Z_i = c_i S_i / T_i$ [7], from which we can see that Z_i is only related to the amount of dataset S_i when c_i and T_i are given.

C. Communication Resource Consumption Model

1) Uplink: The transmission time for transmitting the local model $w_i^r(t)$ is denoted by $T_{\text{up}}^{i,r}$. Since the dimensions of local models are fixed for all UEs that participate in local training, the data size of the local model on each UE is constant and is denoted by s [7]. The transmission rate of UE i on the wireless channel to the BS at rth communication round is represented by $R_{\text{up}}^{i,r}$. Therefore, we have $\frac{s}{R_{\text{up}}^{i,r}} = T_{\text{up}}^{i,r}$, where $R_{\text{up}}^{i,r} = b_{\text{up}}^{i,r} log_2(1 + SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2))$. Specifically, $b_{\text{up}}^{i,r}$ is the bandwidth needed for transmitting the local model of UE i. In addition, $SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2) = \frac{P_{\text{up}} G(D_1)}{\sum_{j=1}^{N_I} PG(D_2^{(j)}) + \delta^2}$ is the signal-to-interference-plus-noise-ratio (SINR). D_1 represents the distance between the UE and BS, $\mathbf{D}_2 = [D_2^{(1)}, D_2^{(2)}, ...D_2^{(j)}, ..., D_2^{(N_I)}]$ is the distance vector for all interfering UEs of UE i, N_I is the number of interfering UEs with $N_I \leq N$, δ^2 is the noise power, P_{up} is the transmit power of the UE, and $G(\cdot)$ is the wireless channel gain between the BS and UE. Furthermore, let β_{up} be the SINR threshold that the BS can successfully decode the received updates from UE i. Therefore, local model transmission is successful only if $SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2) > \beta_{\text{up}}$.

2) DownLink: The transmission time for transmitting the global model g_r is denoted by $T_{\text{down}}^{i,r}$. From Section II. A, we see that the dimensions of the global model g_r is similar to that of each UE's model. Therefore, the data size of the global model is also equal to s [8]. We assume that the transmission rate of the BS at rth communication round is represented by $R_{\text{down}}^{i,r}$. Therefore, we have $\frac{s}{R_{\text{down}}^{i,r}} = T_{\text{down}}^{i,r}$, where

$R^{i,r}_{\text{down}} = b^{i,r}_{\text{down}} log_2(1+SNR_{\text{down}}(D_1))$. Specifically, $b^{i,r}_{\text{down}}$ is the consumed bandwidth for transmitting the global model g_r to UEs at the beginning of communication round r. In addition, $SNR_{\text{down}}(D_1) = \frac{P_{\text{down}} G(D_1)}{\delta^2}$, where P_{down} is the transmit power of the BS allocated to all UEs. Furthermore, the global model transmission is successful only if $SNR_{\text{down}}(D_1) > \beta_{\text{down}}$ where β_{down} is the SNR threshold.

III. Communication and Computing Resources Consumed for FL enabled Edge Intelligence

In this section, we theoretically analyze SINR, SNR, as well as wireless bandwidth and computing resources consumed to support FL enabled edge networks.

A. SINR Analysis for Uplink

1) PDF of SINR: As UEs are geographically distributed as homogeneous PPP with intensity λ_i, the number of UEs within the coverage of the BS is a variable of Poisson distribution with density parameter $\pi(r_0)^2\lambda_i$, where r_0 is the radius of the coverage circle. For a specific UE i, the signal power ($S_{\text{up}} = P_{\text{up}} G(D_1)$) is also a random variable, as it only relates to the distance D_1 and P_{up} is always fixed for each UE.

Proposition 1. *The PDF of the distance D_1 between a specific UE and the serving BS is $f_{D_1}(d_1) = 2d_1/r_0^2$.*

The proof of Proposition 1 is similar to that of Proposition 2 in [9]. Therefore, we can obtain the PDF of signal power (*i.e.*, $f_{S_{\text{up}}} = f_{D_1}(d_1)$). Next, we investigate the distribution of the received interference on the uplink. Note that only transmitting UEs located in the interfering area with radius d_0 can contribute to the interference. We assume that the number of UEs within the interfering area is $N_{\mathcal{A}} (N_{\mathcal{A}} \geq N_I)$ which is also a variable of Poisson distribution with density parameter $\pi(d_0)^2\lambda_i$. Moreover, the transmission time for UEs is represented by t_{up}. Therefore, the transmitting UEs during $[-t_{\text{up}}, t_{\text{up}}]$ can contribute to interference. For a specific UE, the number of interfering UEs is distributed as PPP with parameter $2t_{\text{up}}\lambda_a$ where λ_a is the arrival rate of interfering UEs. Therefore, the interference probability of a transmitting UE during $[-t_{\text{up}}, t_{\text{up}}]$ is $\Pr(active) = 1 - \exp\{-2t_{\text{up}}\lambda_a\}$.

Therefore, the probability of the number of interfering UEs $N_I = n_I$ given $N_{\mathcal{A}} = n_0$ is $\Pr(N_I = n_I | N_{\mathcal{A}} = n_0) = C^{n_I}_{n_0}(1-\exp\{1-2t_{\text{up}}\lambda_a\})^{n_I} \cdot (\exp\{1-2t_{\text{up}}\lambda_a\})^{n_0-n_I}$, where $C^{n_I}_{n_0}$ is the combination number. Therefore, the PDF of N_I is

$$f_{N_I}(n_I) = \sum_{n_0=n_I}^{N} \Pr(N_I = n_I | N_{\mathcal{A}} = n_0) \Pr(N_{\mathcal{A}} = n_0), \quad (1)$$

where $\Pr(N_{\mathcal{A}} = n_0) = \frac{(\pi(d_0)^2\lambda_i)^{n_0}}{n_0!} \exp\{-\pi(d_0)^2\lambda_i\}$. Based on Proposition 1, we can rationally express the PDF of interference I_i generated by UE i as $f_{I_i}(I_i = P_{\text{up}} G(d_2^{(i)})) = f_{D_2^{(i)}}(d_2^{(i)}) = 2d_2^{(i)}/d_0^2$. As the total interference $I(N_I, \mathbf{D}_2)$ is effected by the number of interfering UEs N_I and the distance of these interfering UEs $\mathbf{D}_2$, we have the PDF of $I(N_I, \mathbf{D}_2)$, *i.e.*, $f_I(N_I = n_I, \mathbf{D}_2 = \mathbf{d}_2) = f_{N_I}(n_I) \left(\frac{2}{(d_0)^2}\right)^{n_I} \prod_{n=1}^{n_I} d_2^{(n)}$. Therefore, the PDF of SINR can be given by $f_{SINR_{\text{up}}} = f_{D_1}(d_1) f_I(N_I = n_I, \mathbf{D}_2 = \mathbf{d}_2)$.

2) Transmission Success Rate: For the distance between the UE and the BS, intuitively $f_{SINR_{\text{up}}} < \beta_{\text{up}}$ when $D_1 > d_0$ [9]. Therefore, the satisfying range of D_1 is $(0, d_0]$. Therefore, when given $D_1 = d_1$, we can obtain the number of interfering UEs N_I and the location of these interfering UEs. Let $\overline{n}_I$ represent the mean of random variable N_I. Based on the UE distribution and interfering UE arrival models, we can derive $\overline{n}_I \triangleq E(N_{\mathcal{A}})\Pr(active) = \pi(d_0)^2\lambda_i \left(1 - \exp\{1 - 2t_{\text{up}}\lambda_I\}\right)$, where $E(N_{\mathcal{A}})$ represents the mean of the number of UEs located at the area $\mathcal{A}$ of $(D_1, N_I, \mathbf{D}_2)$ that satisfies $SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2) > \beta_{\text{up}}$.

Therefore, SINR is only related to D_1 and $\mathbf{D}_2$, expressed as $SINR_{\text{up}}(D_1, \mathbf{D}_2) = \frac{P_{\text{up}} G(D_1)}{\sum_{i=1}^{\overline{n}_I} I_i + \delta^2}$, where $I_i = P_{\text{up}} G(D_2^{(i)})$ represents the interference generated by UE i. Therefore, we have $\Pr(SINR_{\text{up}} > \beta_{\text{up}}) = \Pr\left(\sum_{i=1}^{\overline{n}_I} I_i < \frac{P_{\text{up}} G(D_1)}{\beta_{\text{up}}} - \delta^2\right)$, where $\sum_{i=1}^{\overline{n}_I} I_i$ follows a *normal distribution* $N(\mu_I, \sigma_I^2)$ as the number of UEs involved in local model training is large enough. Furthermore, we have $\mu_I = \overline{n}_I E(I_i)$ and $\sigma_I = \sqrt{\overline{n}_I} D(I_i)$, which are the mean and variance of I_i respectively, where $d_{\min}$ is the minimum distance between UEs and the BS, and we define $G(\cdot) = G_1'$ and $G_1(\cdot) = G_2'(\cdot)$. Similarly, we can obtain $\Pr(SNR_{down} > \beta_{\text{down}}) = \Pr(\frac{P_{\text{down}} G(D_1)}{\delta^2} > \beta_{\text{down}}) = \int_{d_1=d'_{\min}}^{r_0} f_{D_1}(d_1) d(d_1)$ when we $G(\cdot)$ monotonically increases.

Let $Y = \frac{I - \mu_I}{\sigma_I}$, where $I = \sum_{i=1}^{\overline{n}_I} I_i$ and $I \sim N(\mu_I, \sigma_I)$. Therefore, we have $Y \sim N(0,1)$, where $\Pr\left(\sum_{i=1}^{\overline{n}_I} I_i < \frac{P_{\text{up}} G(d_1)}{\beta_{\text{up}}} - \delta^2\right) = \Phi(\xi(d_1))$, where Φ is the cumulative distribution function (CDF) of standard normal distribution and $\xi(d_1) = \frac{1}{\sigma_I}(\frac{P_{\text{up}} G(d_1)}{\beta_{\text{up}}} - \delta^2 - \mu_I)$. Therefore, we have

$$\Pr\left(SINR_{\text{up}} > \beta_{\text{up}}\right) = \int_{d_1=d_{\min}}^{d_0} f_{D_1}(d_1) \Phi(\xi(d_1)) d(d_1). \quad (2)$$

B. Wireless Bandwidth Consumed for Transmitting Models

According to Section II. C, the bandwidth consumed for transmitting the local model $w_i^r(t)$ at rth communication round is given by $b^{i,r}_{\text{up}} = \frac{s}{T^{i,r}_{\text{up}} \log_2(1+SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2)}$. As s and $T^{i,r}_{\text{up}}$ are constant, the PDF of $b^{i,r}_{\text{up}}$ for UE i is equal to $f_{SINR_{\text{up}}}$. Therefore, the mean of bandwidth for all UEs transmitting local models during K communication rounds is $\overline{B}_{\text{up}} = K \cdot \sum_{i=1}^{n} b^{i,r}_{\text{up}} f_{SINR_{\text{up}}}$. Similarly, the bandwidth for transmitting the global models during K communication rounds is given by $\overline{B}_{\text{down}} = K \cdot \sum_{i=1}^{n} b^{i,r}_{\text{down}} f_{SNR_{\text{down}}}$.

C. Computing Resources Consumption

Obviously, the total computing resources needed to support local model training are affected by the amount of training data and the number of training UEs. We assume that the amount of data samples among UEs follows the *normal distribution*, *i.e.*, $\mathcal{S}_i \sim N(\mu_i, \sigma_i^2)$. Note that μ_i or/and σ_i^2 could be different for

specific UEs. Therefore, as the computing resources consumed of UE i for one local iteration is $Z_i = c_i S_i / T_i$, the PDF of Z_i is equal to $f_{S_i}(s_i)$.

For a specific UE i, if $SNR_{down} > \beta_{\text{down}}$, we say UE i can successfully receive the global model. In other words, UE i will continue to perform local training in the next communication round and consume certain computing resources. Let $\hat{\boldsymbol{Z}} = \{\hat{Z}_1, ..., \hat{Z}_n\}$ indicate the certain computing resources consumed by UEs, where the value of $\hat{Z}_i$ is set to z_i if $SNR_{down} > \beta_{\text{down}}$ and 0 otherwise. Therefore, we can obtain the PDF of $\hat{Z}_i$ as $f_{\hat{Z}_i} = f_{Z_i}(z_i)\Pr(SNR_{down} > \beta_{\text{down}})$.

In FL, the consumption of computing resources on each UE is independent as UEs train local model independently. Therefore, we can derive the mean of computing resources consumed by all UEs for one local iteration as $\overline{C}_{UE} = \sum_{i=1}^{n} z_i f_{\hat{Z}_i}(\hat{Z}_i = z_i)$. Therefore, the total computing resources consumed is given by $C_{\text{total}} = \tau K \overline{C}_{UE}$.

IV. The Relationship between FL Performance and Consumed Resources

Indeed, both the unsuccessful transmission of the local and global model affect the aggregation and the updating of FL tasks. Therefore, we need to analyze how the computing and communication resources consumed affect the FL performance by evaluating both the local and global model accuracy.

A. Local Model Accuracy

Practically, similar to that in [7], each UE solves the local optimization problem

$$\min_{h_i \in \mathbb{R}^d} G_i(g_r, h_i) \triangleq F_i(g_r + h_i) - (\nabla F_i(g_r) - \zeta \nabla F(g_r))^{\mathrm{T}} h_i, \tag{3}$$

where ζ is constant and h_i represents the difference between the global model and the local model for UE i. In this work, we use the GD method, *i.e.*, $h_i^{(r)(t+1)} = h_i^{(r)(t)} - \xi \nabla G_i(g_r, h_i^{(r)(t)})$, where ξ is the step size and $h_i^{(r)(t)}$ is the value of h_i. Moreover, $\nabla G_i(g_r, h_i^{(r)(t)})$ is the gradient of $G_i(g_r, h_i)$. $g_r + h_i^{(r)(t)}$ is the local model of UE i at local iteration t. For a small step ξ, we can derive a set of solutions $h_i^{(r)(0)}, \cdots, h_i^{(r)(\tau)}$, which satisfies $G_i(g_r, h_i^{(r)(0)}) \geq \cdots \geq G_i(g_r, h_i^{(r)(\tau)})$.

To provide the convergence condition for GD method, we introduce local model accuracy loss ϵ_l [7], *i.e.*, $G_i(g_r, h_i^{(r)(t)}) - G_i(g_r, h_i^{(r)*}) \leq \epsilon_l (G_i(g_r, h_i^{(r)(0)}) - G_i(g_r, h_i^{(r)*}))$, where the local model accuracy is $1-\epsilon_l$ and $h_i^{(r)*}$ represents the optimal solution. To achieve the local and global model accuracy given in the next subsection, we first make the following three assumptions on the loss function $F_i(w)$, as that in [7], [8],

- Assumption 1: Function $F_i(w)$ is L-Lipschitz, *i.e.*, $\forall w, w' \in \mathbb{R}^d, \| \nabla F_i(w) - \nabla F_i(w') \| \leq L \| w - w' \|$.
- Assumption 2: Function $F_i(w)$ is γ-strongly convex, *i.e.*, $\forall w, w' \in \mathbb{R}^d, F_i(w) \geq F_i(w') + \langle \nabla F_i(w'), (w - w') \rangle + \frac{\gamma}{2} \| w - w' \|^2$.
- Assumption 3: $F_i(w)$ is twice-continuously differentiable. And $\gamma I \leq \nabla^2 F_i(w) \leq LI$.

Based on the assumptions, we can obtain the lower bound on the number of local iterations, shown as Proposition 2. The proof of Proposition 2 is similar to that in Appendix A in [7]. The lower bound reflects the growing trend of the number of local iterations with respect to local model accuracy, which can approximate the consumption of computing resources for training local models.

Proposition 2. *Local model accuracy loss ϵ_l is achieved if $\xi < \frac{2}{L}$ and run the GD method $\tau \geq \lceil \frac{2}{(2-L\xi)\xi\gamma} \ln \frac{1}{\epsilon_l} \rceil$ iterations during each communication round at each UE that participants in local training.*

B. Global Model Accuracy

Let $\hat{S} = \{\hat{S}_1, \hat{S}_2, ..., \hat{S}_n\}$ indicate whether local models successfully contribute to the global aggregation when a UEs sends its local model to the BS, where the value of $\hat{S}_i$ is set to S_i if $SINR_{\text{up}}(D_1, N_I, \mathbf{D}_2) > \beta_{\text{up}}$ and 0 otherwise. Therefore, the probability of $\hat{S}$ is given by

$$\begin{aligned} &\Pr\left(\hat{S} = \{\hat{S}_1, \hat{S}_2, ..., \hat{S}_n\}\right) = \\ &\left(\Pr\left(SINR_{\text{up}} > \beta_{\text{up}}\right)\right)^{n_s} \left(1 - \Pr\left(SINR_{\text{up}} > \beta_{\text{up}}\right)\right)^{n - n_s}, \end{aligned} \tag{4}$$

where n_s is the mean of the number of UEs that successfully send the local models to the BS. Therefore, the global model at r-th communication round can be rewritten as $g_r = \frac{\sum_{i=1}^{n} \hat{S}_i w_i^r(t)}{\sum_{i=1}^{n} \hat{S}_i}$, from which we can analysis the impact of the SINR on the global model.

In FL algorithm, a global model accuracy is also needed. For a specific FL task, we define ϵ_g as its global model accuracy loss, *i.e.*, $F(g_r(\hat{\mathcal{S}}, SINR_{\text{up}})) - F(g^*) \leq \epsilon_g (F(g_0) - F(g^*))$, where g^* is the actual optimal solution. Moreover, we provide the following Proposition 3 about the number of communication rounds for achieving global model accuracy $1 - \epsilon_g$. The proof of Proposition 3 is similar to that in Appendix B in [7].

Proposition 3. *Global model accuracy $1-\epsilon_g$ is achieved if the number of communication rounds K meets $K \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{(1-\epsilon_l)\gamma^2 \zeta} \rceil$ when running FL algorithm shown as Algorithm 1 with $0 < \zeta < \frac{\gamma}{L}$*

V. Discussions of Three Different cases

In this section, we discuss three special cases and derive the explicit expression of the model accuracy under FL framework as a function of the amount of consumed computing resources and communication resources based on the sufficiency of respective communication and computing resources.

A. Sufficient Communication and Computing Resources

When both communication and computing resources are sufficient, we can approximate the communication and computing resources needed for the FL task based on Proposition 2 and Proposition 3. Specifically, the communication resources needed for transmitting local models should meet $\overline{B}_{\text{up}} =$

$K \cdot \sum_{i=1}^{n} b_{\text{up}}^{i,r} f_{SINR_{\text{up}}} \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{(1-\epsilon_l)\gamma^2\zeta} \rceil \cdot \sum_{i=1}^{n} b_{\text{up}}^{i,r} f_{SINR_{\text{up}}}$. Similarly, we can obtain the communication resources needed for transmitting the global model, shown as $\overline{B}_{\text{down}} \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{(1-\epsilon_l)\gamma^2\zeta} \rceil \cdot \sum_{i=1}^{n} b_{\text{down}}^{i,r} f_{SNR_{\text{down}}}$. Furthermore, given local accuracy ϵ_l with $\xi < \frac{2}{L}$, the total computing resources needed should meet the following constraint, $C_{\text{total}} \geq \overline{C}_{\text{UE}} \cdot \lceil \frac{2}{(2-L\xi)\xi\gamma} \ln \frac{1}{\epsilon_l} \rceil \cdot \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{(1-\epsilon_l)\gamma^2\zeta} \rceil$.

B. Sufficient Computing Resources and Insufficient Communication Resources

As computing resources are sufficient, the number of local iterations still follows Proposition 2. However, Proposition 3 may not be met due to the lack of communication resources. As a result, the number of communication rounds K will decrease even cannot achieve the required global accuracy. In this case, the maximal number of communication rounds $K_{\max}$ is given by $K_{\max} = \lfloor \min\{\frac{B_{\text{down}}^{\max}}{\overline{B}_{\text{down}}^{i}}, \frac{B_{\text{up}}^{\max}}{\overline{B}_{\text{up}}^{i}}\} \rfloor$, where $B_{\text{up}}^{\max}$ and $B_{\text{down}}^{\max}$ are the maximal available bandwidth that can be used for FL on the uplink and downlink respectively. To achieve the required global accuracy even the communication round is limited, we can reasonably expect the real achieved global model accuracy loss $\widetilde{\epsilon}_g$ can be expressed by $\widetilde{\epsilon}_g = \exp\left(-K\left(\frac{(1-\widetilde{\epsilon}_l)\gamma^2\zeta}{2L^2}\right)\right)$ [7]. Therefore, we have $K = \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{(1-\widetilde{\epsilon}_l)\gamma^2\zeta} \rceil$, where $\widetilde{\epsilon}_l$ is the realistic local model accuracy loss. In addition, the number of communication rounds K should meet $K \leq K_{\max}$. Therefore, we have $\widetilde{\epsilon}_l \leq \lfloor 1 - \frac{2L^2 \ln \frac{1}{\epsilon_g}}{K_{\max}\gamma^2\zeta} \rfloor$.

Furthermore, based on Proposition 2, the number of local iterations $\tau \geq \lceil \frac{2}{(2-L\xi)\xi\gamma} \ln \frac{K_{\max}\gamma^2\zeta}{K_{\max}\gamma^2\zeta - 2L^2 \ln \frac{1}{\widetilde{\epsilon}_g}} \rceil$. Therefore, the total amount of computing resources consumed is given by $C_{\text{total}} \geq \overline{C}_{\text{UE}} \cdot \lceil \frac{2}{(2-L\xi)\xi\gamma} \ln \frac{K_{\max}\gamma^2\zeta}{K_{\max}\gamma^2\zeta - 2L^2 \ln \frac{1}{\widetilde{\epsilon}_g}} \rceil \cdot \lceil \frac{2L^2 \ln \frac{1}{\widetilde{\epsilon}_g}}{(1-\widetilde{\epsilon}_l)\gamma^2\zeta} \rceil$.

C. Sufficient Communication Resources and Insufficient Computing Resources

Similarly, the number of local iterations should meet $\tau \cdot \overline{C}_{\text{UE}} \leq \sum_{i=1}^{n} C_i$ where C_i represents the maximal computing resources used for local training on UE i. To achieve the required local accuracy though the local iterations are limited, we can reasonably expect that the real local global model accuracy loss $\widetilde{\epsilon}_l$ is expressed by $\widetilde{\epsilon}_l = \exp\left(-\tau\frac{(2-L\xi)\xi\gamma}{2}\right)$ [7].

Therefore, when $\xi < \frac{2}{L}$, we have $\widetilde{\epsilon}_l \geq \exp\left(\frac{(L\xi-2)\xi\gamma \sum_{i=1}^{n} C_i}{2\overline{C}_{\text{UE}}}\right)$. Moreover, we can derive the lower bound of the number of communication rounds as $K \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{\left(1-\exp\left(\frac{(L\xi-2)\xi\gamma \sum_{i=1}^{n} C_i}{2\overline{C}_{\text{UE}}}\right)\right)\gamma^2\zeta} \rceil$.

Therefore, the bandwidth for transmitting local models and the global model are respectively given by $\overline{B}_{\text{up}} \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{\left(1-\exp\left(\frac{(L\xi-2)\xi\gamma \sum_{i=1}^{n} C_i}{2\overline{C}_{\text{UE}}}\right)\right)\gamma^2\zeta} \rceil \sum_{i=1}^{n} b_{\text{up}}^{i,r} f_{SINR_{\text{up}}}$ and $\overline{B}_{\text{down}} \geq \lceil \frac{2L^2 \ln \frac{1}{\epsilon_g}}{\left(1-\exp\left(\frac{(L\xi-2)\xi\gamma \sum_{i=1}^{n} C_i}{2\overline{C}_{\text{UE}}}\right)\right)\gamma^2\zeta} \rceil \sum_{i=1}^{n} b_{\text{down}}^{i,r} f_{SNR_{\text{down}}}$.

VI. Simulations

We consider an FL enabled edge network composed of multiple UEs and one central BS with a cloud server serving as the FL aggregator. The coverage of the BS is a circular area with a radius of 1KM. The radius of the interfering area is set to 200m. The transmit power of UEs and the serving BS is set to 20dBm and 43dBm respectively [9]. Moreover, the noise power is set to -173dBm [9]. The density of interfering UEs λ_a is set to $1\text{UE}/\text{m}^2$. The path loss is modeled as $g(D_1) = 34 + 40log(D_1)$ [5]. The size of transmission model s is set to 28.1kbits [7]. The number of CPU cycles required for computing one sample data is randomly chosen within $[1, 4] \cdot 10^4$ cycles/sample [7]. In addition, we consider the multi-class classification problem over MINIST datasets where datasets of UEs are splitted randomly with 75% and 25% for training and testing. Moreover, we use a two-layer fully connected neural network, where the activation function is ReLU. The learning rate is 0.03.

Algorithm 1 : FL Algorithm.

Input: ϵ_l, ϵ_g.
output: g_r, τ, K.

```
Initialization: w_i^1(0) = 0, g_1 = 0.
for r = 1, 2, ... do
    Each UE calculates ∇F_i(g_r) and sends it to BS
    The BS calculates ∇F(g_r) and broadcasts it to UEs
    Parallel Each UE i = 1, 2, ..., n
    Initialization: t = 0, h_i^{(r)(0)} = 0.
    Repeat
    Every V steps set h_i^{(r)*} = h_i^{(r)(t)}.
    Update h_i^{(r)(t+1)} = h_i^{(r)(t)} − ξ∇G_i(g_r, h_i^{(r)(t)}).
    Set w_i^r(t) = g_r + h_i^{(r)(t)}.
    if (G_i(g_r, h_i^{(r)(t)}) − G_i(g_r, h_i^{(r)*})) / (G_i(g_r, h_i^{(r)(0)}) − G_i(g_r, h_i^{(r)*})) > ε_l then
        Set t = t + 1
    else
        Each UE i sends w_i^r(t) to the BS.
    end if
    The BS calculates g_r and sends it to UEs
    if (F(g_r(Ŝ, SINR_up)) − F(g*)) / (F(g_0) − F(g*)) < ε_g then
        Break;
    end if
end for
Set τ = t, K = r.
```

First, we examine the local and global model transmission success rates with varying UE density. Fig. 3 shows $\Pr(\text{SINR}_{\text{up}} > \beta_{\text{up}})$ on uplink (UL) and $\Pr(\text{SNR}_{\text{down}} > \beta_{\text{down}})$ on downlink (DL). From Fig.3, we can see that the curves of analytical results match closely to simulations for both the uplink and downlink. As expected, under both $\beta_{\text{up}} = -15$dB and $\beta_{\text{up}} = -12$dB scenarios, $\Pr(\text{SINR}_{\text{up}} > \beta_{\text{up}})$ decreases with the UE density. This is because that the interference increases with the UE density. In addition, we also find that

$\Pr(\text{SINR}_{\text{up}} > \beta_{\text{up}})$ under $\beta_{\text{up}} = -15$dB is more than that under $\beta_{\text{up}} = -12$dB due to the more stringent SINR requirement.

After that, we examine the bandwidth consumption in the uplink and the downlink respectively for both analytical and simulation results with respect to the global accuracy loss ϵ_g. Fig. 4 and Fig. 5 show the bandwidth consumption in the uplink and the downlink changes with the global accuracy loss respectively, where both the bandwidth consumption in the uplink and the downlink decrease with the global accuracy loss. In addition, we also find that the lower local accuracy leads to more bandwidth consumption to guarantee a specific global accuracy when training i.i.d data. The reason is that the lower local accuracy needs more communication rounds to aggregate the local models to achieve a certain global accuracy, and thus consumes more bandwidth.

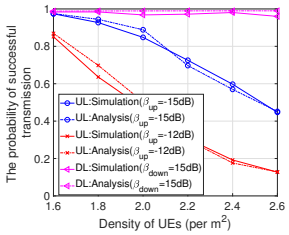

Fig. 3: Comparison of successful transmission probability.

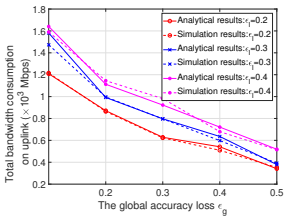

Fig. 4: Comparison of bandwidth on the uplink.

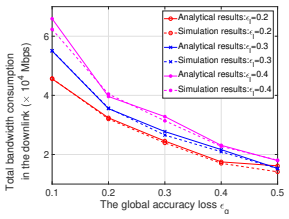

Fig. 5: Comparison of bandwidth on the downlink.

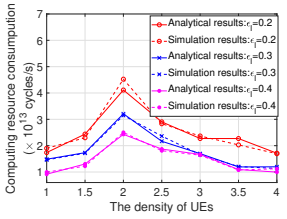

Fig. 6: Comparison of the computing resources.

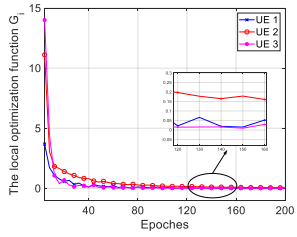

Fig. 7: Local training during each communication round.

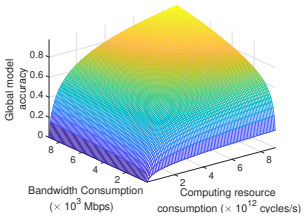

Fig. 8: Trade-off between computing resources and bandwidth.

In the following, we examine the computing resource consumption for both analytical and simulation results with the density of UEs. Fig. 6 shows the computing resource consumption changes with the density of UEs. From Fig. 6, we can see that the computing resource consumption increases in the beginning and then decreases with the density of UEs. This is because that the number of UEs that participate in local training increases in the beginning and then decreases due to the impact of SNR.

Next, we verify the convergence property by observing the local optimization function whether converges with the local training. As shown in Fig. 7, the local optimization function convergences with the number of local trainings increasing. After that, we aim to verify the relationship between the consumption of computing resources and communication resources. As shown in Fig. 8, when we fix the global model accuracy, we can reduce the bandwidth consumption by increasing the computing resource consumption and we can also reduce the computing resource consumption by increasing the bandwidth consumption. Moreover, we can also see, when the amount of available computing resources/bandwidth is vitally small, it makes little sense to improve the global model accuracy no matter how many bandwidth (computing resources) we increase.

VII. Conclusion

Wireless edge network intelligence enabled by FL has been widely acknowledged as a very promising means to address a wide range of challenging network issues. In this paper, we have theoretically analyzed how accurate of an ML model can be achieved by using FL and how many resources are consumed to guarantee a certain local/global accuracy. Specifically, we have derived the explicit expression of the model accuracy under FL framework, as a function of the amount of computing/communication resources for FL empowered wireless edge networks. Numerical results validate the effectiveness of our theoretical modeling. The modeling and results can provide some fundamental understanding for the trade-off between the learning performance and consumed resources, which is useful for promoting FL empowered wireless network edge intelligence.

References

[1] B. Jovanović, "Internet of things statistics for 2021 – taking things apart," https://dataprot.net/statistics/iot-statistics/.

[2] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[4] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[5] Y. Liu, G. Feng, Y. Sun, S. Qin, and Y.-C. Liang, "Device association for ran slicing based on hybrid federated deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, 2020.

[6] C. Hennig and M. Kutlukaya, "Some thoughts about the design of loss functions," *REVSTAT–Statistical Journal*, vol. 5, no. 1, pp. 19–39, 2007.

[7] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, 2020.

[8] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2021.

[9] Y. Sun, L. Zhang, G. Feng, B. Yang, B. Cao, and M. A. Imran, "Blockchain-enabled wireless internet of things: Performance analysis and optimal communication node deployment," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5791–5802, 2019.