Energy Harvesting Aware Client Selection for Over-the-Air Federated Learning

Caijuan Chen*, Yi-Han Chiang†, Hai Lin†, John C.S. Lui‡, and Yusheng Ji*
*The Graduate University for Advanced Studies (SOKENDAI), and National Institute of Informatics (NII), Tokyo, Japan
†Department of Electrical and Electronic Systems Engineering, Osaka Metropolitan University, Osaka, Japan
‡Department of Computer Science Engineering, The Chinese University of Hong Kong, Hong Kong
Email: chencaijuan@nii.ac.jp, chiang@omu.ac.jp, lin@omu.ac.jp, cslui@cse.cuhk.edu.hk, kei@nii.ac.jp

***Abstract*—Federated learning (FL) has been widely regarded as a promising distributed machine learning technology that utilizes on-device computation while protecting clients' data privacy. To adapt FL to wireless networks, the over-the-air (OTA) computation, which employs the superposition nature of wireless waveforms, can prevent excessive consumption of the communication resources. However, energy harvesting technology can overcome the energy limitation of clients to realize durable computation. Despite the existing works devoted to OTA FL from various aspects, they mostly neglect jointly performing client selection and energy management for energy harvesting devices. In this paper, we investigate the combined problem of client selection and energy management for OTA FL and formulate it as a nonlinear integer programming (NIP) problem to minimize the optimality gap. To solve the NIP problem, we propose a client selection scheme that jointly considers channel state information, residual battery capacities, and dataset size. Our simulation results show that the proposed solution outperforms other comparison schemes within various parameter settings.**

***Index Terms*—Federated learning, over-the-air computation, client selection, energy harvesting.**

I. Introduction

With the increasing number of deployed Internet of Things (IoT) devices, a vast amount of data is constantly being generated by these devices. Developers often use deep learning methods to extract valuable information from this data. Furthermore, due to advancements in hardware, the computing and storage capabilities of terminal devices, such as mobile phones, smartwatches, and wearable devices, have been greatly improved. Federated learning (FL) [1] is regarded as a potential solution to performing information extraction on the distributed IoT data and preventing privacy leaks.

Under the FL paradigm, clients cooperatively train a shared model using their local data and upload the updates to the parameter server (PS). Then, the PS conducts the updated aggregation and broadcasts the global model updates to clients. The clients need to communicate with the PS multiple rounds during the training process. However, since communication resources are usually limited, the communication bottleneck is a dominant issue that needs to be addressed. Different client selection [2]–[4] strategies have been proposed to realize communication efficiency. An energy-efficient client selection is designed in [2] to realize green computation. In [4], client selection and bandwidth allocation are jointly optimized to realize fast convergence.

The above works related to FL are implemented with digital transmission. Recently, works [5], [6] based on over-the-air (OTA) computation have been successfully used to reduce the communication cost of FL. Compared with digital transmission, clients can share the same wireless channel via analog transmission due to the superposition property of the multiple access channel. Recent research works related to OTA FL systems focus on power control [7], data heterogeneity [8], and energy constraints [9]. The authors in [7] design a client selection scheme with transmission power control and carry out a convergence analysis for OTA FL. Convergence analysis for heterogeneous data is given in [8], and the conclusion is that FL based on analog transmission could achieve convergence even when heterogeneous data and fading channels existed. The Lyapunov optimization technique is used in [9] to design a stable client selection scheme under the energy constraint via an online optimization formulation.

However, the majority of existing works related to OTA FL do not consider energy harvesting, which is a technology that enables clients to acquire energy resources from the environment, such as wind, solar power, and human motion, allowing clients to continuously engage in training and realize green computing. Some prior works combine energy harvesting with FL via digital transmission [10], [11]. In [10], clients are powered by separate energy harvesting resources, and an online client selection and client association algorithm is proposed to minimize the training loss within multiple base station settings. The authors in [11] investigate two different energy harvesting settings: deterministic energy arrivals and stochastic energy arrivals. They conclude that energy harvesting makes sense for sustainable distributed learning.

In this paper, we study the joint problem of client selection and energy management for an OTA FL system. As a result, we first formulate it as a nonlinear integer programming (NIP)

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001136

problem with an objective function to minimize the optimality gap over multiple training rounds. Due to the intractability of the optimality gap, we provide a thorough convergence analysis, based on which we transform the NIP problem into an online NIP problem, so that we can optimize the client selection in a per-training round basis. To achieve this, we first define the channel-energy-data (CED) coefficient to quantitatively characterize the influence of selecting one certain client on the convergence rate, and then propose a client selection algorithm that jointly considers channel state information, residual battery capacity, and dataset size. Our simulation results demonstrate that the proposed solution outperforms other comparison schemes due to wisely selecting clients according to the CED coefficient. In addition, the proposed solution remains superior even if the impacts of noisy channels become pronounced.

II. System Model and Problem Formulation

We consider an OTA FL system that consists of K single-antenna clients and one single-antenna PS. The set of all clients is denoted as $\mathcal{K}$. Each client $k \in \mathcal{K}$ has a number of feature-label pairs, which can be denoted as $\mathcal{D}_k = \{(\boldsymbol{x}_{k,i}, y_{k,i}), i \in \{1, \ldots, D_k\}\}$, where $\boldsymbol{x}_{k,i}$ is the feature vector, $y_{k,i}$ is the ground-truth label, and D_k is the total number of local training samples owned by client k. The total number of samples for all clients is $D = \sum_{k\in\mathcal{K}} D_k$. Clients can harvest energy from various energy harvesting resources. In this section, we first provide some background knowledge about the FL model, the communication model, the energy consumption and harvesting model. Then, we formulate the problem to minimize the optimality gap by jointly optimizing the client selection and energy management for the OTA FL system.

A. Federated Learning Model

In the considered OTA FL system, the PS continually exchanges information with the clients over multiple training rounds until the training process converges. At the t-th training round, it carries out the following four steps:

- *Client selection*: Initially, the PS selects a subset of clients to participate in the training. The client selection vector is defined as $\boldsymbol{\beta}_t = [\beta_{1,t}, \beta_{2,t}, \ldots, \beta_{k,t}]$, where $\beta_{k,t} \in \{0, 1\}$ indicates whether client k is selected, with 1 indicating that client k is selected to upload the local gradients to the PS, and 0 otherwise.
- *Global model broadcast*: The global model $\mathbf{w}_t = [\mathrm{w}_t^1, \mathrm{w}_t^2, \ldots, \mathrm{w}_t^s] \in \mathbb{R}^s$ is broadcasted to the selected clients, where s denotes the total number of weight parameters.
- *Local model update*: After receiving the global model $\mathbf{w}_t$, the local model $\mathbf{w}_{k,t}$ for client k is updated as $\mathbf{w}_{k,t} = \mathbf{w}_t$. The local loss for client k is calculated as $F_{k,t}(\mathbf{w}_{k,t}, \mathcal{D}_k) = \frac{1}{D_k}\sum_{i=1}^{D_k} f(\mathbf{w}_{k,t}, \boldsymbol{x}_{k,i}, y_{k,i})$, where $f(\mathbf{w}_{k,t}, \boldsymbol{x}_{k,i}, y_{k,i})$ is the loss value for the local training sample $(\boldsymbol{x}_{k,i}, y_{k,i})$, $i \in \{1, \ldots, D_k\}$. The local gradient $\mathbf{g}_{k,t} = [\mathrm{g}_{k,t}^1, \mathrm{g}_{k,t}^2, \ldots, \mathrm{g}_{k,t}^s] \in \mathbb{R}^s$ is computed as $\mathbf{g}_{k,t} = \frac{1}{D_k}\sum_{i=1}^{D_k} \nabla f(\mathbf{w}_{k,t}, \boldsymbol{x}_{k,i}, y_{k,i})$, where $\nabla f(\mathbf{w}_{k,t}, \boldsymbol{x}_{k,i}, y_{k,i})$ is the gradient for the local training sample $(\boldsymbol{x}_{k,i}, y_{k,i})$. The learning rate is denoted as η, and the local model update can be computed as $\mathbf{w}_{k,t+1} = \mathbf{w}_{k,t} - \eta\mathbf{g}_{k,t}$. The selected clients upload the local gradients $\mathbf{g}_{k,t}$ to the PS.
- *Model aggregation*: The global gradient is aggregated as

$$\mathbf{g}_t = \frac{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k \mathbf{g}_{k,t}}{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k}. \tag{1}$$

The global model is then updated through the following:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{g}_t. \tag{2}$$

The total loss for all training samples belonging to the selected clients can be expressed as follows:

$$F_t(\mathbf{w}_t, \boldsymbol{\beta}_t) = \frac{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k F_{k,t}(\mathbf{w}_{k,t}, \mathcal{D}_k)}{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k}. \tag{3}$$

B. Communication Model

For the OTA FL system, clients need to upload their gradients to the PS with the shared wireless channel synchronously. In this work, we assume that the channel coefficients are quasi-static during the same training round, but they may fluctuate in different training rounds. Let $\hat{h}_{k,t} \in \mathbb{C}$ represent the channel coefficient of client k at the t-th training round, and $h_{k,t}$ means the magnitude of $\hat{h}_{k,t}$. Let $p_{k,t}$ denote the power control factor of client k at the t-th training round. The power control factor $p_{k,t}$ needs to satisfy the requirement for the selected client k at the t-th training round [12] as follows:

$$p_{k,t} = \frac{\alpha_t D_k}{h_{k,t}}, \tag{4}$$

where α_t is the scaling factor for obtaining the transmission energy (i.e., the second term of (8)) at the t-th training round. The preprocessing operation for local gradients $\mathbf{g}_{k,t}$ at client k during the t-th training round is as follows:

$$\boldsymbol{\varphi}_{k,t} = p_{k,t}\beta_{k,t}\mathbf{g}_{k,t}. \tag{5}$$

Let $\boldsymbol{z}_t = [z_t^1, z_t^2, \ldots, z_t^s] \in \mathbb{C}^s$ represent the additive white Gaussian noise (AWGN) vector, and z_t^s follows the distribution $\mathcal{CN}(0, \sigma^2)$. The received gradients vector $\boldsymbol{r}_t \in \mathbb{C}^s$ at the PS side via OTA computation at the t-th training round is calculated as:

$$\boldsymbol{r}_t = \sum_{k=1}^{K} h_{k,t}\boldsymbol{\varphi}_{k,t} + \boldsymbol{z}_t. \tag{6}$$

To obtain the averaged gradients of the OTA FL system, the PS conducts the postprocessing operation for the received aggregated gradients [7] as follows:

$$\bar{\boldsymbol{r}}_t = \frac{\boldsymbol{r}_t}{\alpha_t \sum_{k\in\mathcal{K}} \beta_{k,t} D_k} = \frac{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k \mathbf{g}_{k,t} + \frac{\boldsymbol{z}_t}{\alpha_t}}{\sum_{k\in\mathcal{K}} \beta_{k,t} D_k}. \quad (7)$$

C. Energy Consumption and Harvesting Model

Let c_k denote the computation energy consumption for one training sample on client k. We assume that c_k remains unchanged for client k, but c_k varies across different clients because of client heterogeneity. The data size D_k for client k may also be different across different clients. By combining (4), the total energy consumption $e_{k,t}^{\text{tot}}$ [9] for client k during the t-th training round can be obtained as follows:

$$e_{k,t}^{\text{tot}} = \underbrace{D_k c_k}_{\substack{\text{Computation energy}\\ \text{consumption}}} + \underbrace{\frac{\alpha_t^2 D_k^2}{h_{k,t}^2} \|\mathbf{g}_{k,t}\|_2^2}_{\text{Transmission energy consumption}}. \quad (8)$$

We assume that each client can harvest energy from its resources. The energy harvesting process is built based on a successive energy arrival model. The newly arrived energy at the t-th training round is as follows: $\boldsymbol{e}_t^{\text{arr}} = [e_{1,t}^{\text{arr}}, \ldots, e_{K,t}^{\text{arr}}]$, where $e_{k,t}^{\text{arr}}$ follows a Poisson process with the parameter $\bar{e}_k$. For each client, the harvested energy can be stored in the battery and used for further operations. The current battery energy is denoted as $b_{k,t}$ at the initial time of the t-th training round for client k. The maximum battery capacity for client k is denoted as $B^{\max}$.

For each selected client k at the t-th training round, the total energy consumption cannot exceed the residual battery capacity, which is denoted as follows:

$$e_{k,t}^{\text{tot}} \leq b_{k,t}. \quad (9)$$

The updated battery energy of the client k is less than the maximum capacity $B^{\max}$. The battery level can be expressed as follows:

$$b_{k,t+1} = \min\left\{b_{k,t} - e_{k,t}^{\text{tot}} + e_{k,t}^{\text{arr}}, B^{\max}\right\}, \quad (10)$$

where $b_{k,1} = B_k^1$ is the initial battery energy at client k.

D. Problem Formulation

The total number of training rounds is set as T. Let $\mathbf{w}^*$ indicate the optimal global mode, and let $F(\mathbf{w}^*)$ represent the global loss of the optimal global model. The optimality gap is defined as $\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*)$ after T training rounds. Our goal is to minimize the optimality gap $\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*)$ within the energy constraint via client selection for the OTA FL system. The optimization problem is formulated as follows:

$$\begin{aligned}\boldsymbol{\mathcal{P}}_1: \quad \min_{\boldsymbol{\beta}} \quad & \mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) \\ \text{s.t.} \quad & e_{k,t}^{\text{tot}} \leq b_{k,t}, \qquad \forall k\in\mathcal{K}, 1\leq t\leq T, \\ & \beta_{k,t} \in \{0,1\}, \quad \forall k\in\mathcal{K}, 1\leq t\leq T,\end{aligned}$$

where the first constraint denotes that the total energy consumption at the t-th training round for each selected client k is less than its residual battery capacity, and the second constraint indicates the binary client selection vector at the t-th training round.

III. Convergence Analysis and Problem Transformation

In this section, we present the convergence analysis of the OTA FL system. Note that the energy consumption and the energy harvesting are introduced into the convergence analysis. The derived convergence result can measure the impacts of the channel state information, residual battery capacity and dataset size on the convergence rate. Based on the convergence result, the transformed problems are rebuilt from $\boldsymbol{\mathcal{P}}_1$.

A. Convergence Analysis

According to (2) and (7), the global weight parameters for the OTA FL system are updated as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\bar{\boldsymbol{r}}_t. \quad (11)$$

For the convenience of convergence analysis, we make the following three assumptions on the loss function as used in [4], [7].

Assumption 1: (***l*-Smoothness**) This indicates that a nonnegative constant l exists for parameters $\mathbf{w}$ and $\mathbf{v}$ to make the inequality satisfy the following:

$$F(\mathbf{w}) - F(\mathbf{v}) \leq (\mathbf{w}-\mathbf{v})^T \nabla F(\mathbf{v}) + \frac{l}{2}\|\mathbf{w}-\mathbf{v}\|_2^2. \quad (12)$$

Assumption 2: (**PL Inequality**) For a nonnegative μ constant, the Polyak-Lojasiewicz (PL) condition should hold the following:

$$\|\nabla F(\mathbf{w})\|_2^2 \geq 2\mu\left[F(\mathbf{w}) - F(\mathbf{w}^*)\right]. \quad (13)$$

Assumption 3: (**Gradient Bound**) The local gradients have the bound constraint of the global gradients as follows:

$$\|\nabla f(\mathbf{w})\|_2^2 \leq \lambda_1 + \lambda_2 \|\nabla F(\mathbf{w})\|_2^2, \quad (14)$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

The optimality gap at the t-th training round is defined as $\mathbb{E}[F(\mathbf{w}_{t+1})] - F(\mathbf{w}^*)$. We can derive the relationship of the optimality gap for two adjacent rounds as in Lemma 1.

Lemma 1: Based on Assumptions 1-3, when the learning rate satisfies $\eta = \frac{1}{l}$, given the residual battery capacity vector $\boldsymbol{b}_t$ and the client selection vector $\boldsymbol{\beta}_t$, the optimality gap at the t-th training round within the energy constraint can be represented as follows

$$\mathbb{E}[F(\mathbf{w}_{t+1})] - F(\mathbf{w}^*) \leq \psi_t\left(\mathbb{E}[F(\mathbf{w}_t)] - F(\mathbf{w}^*)\right) + \frac{\lambda_1}{l}\phi_t, \quad (15)$$

where ψ_t and ϕ_t can be expressed as follows:

$$\psi_t = 1 - \frac{\mu}{l} + \frac{2\mu\lambda_2\phi_t}{l}, \tag{16}$$

$$\phi_t = \frac{s\sigma^2}{2\left(\sum_{k\in\mathcal{K}} D_k\beta_{k,t}\right)^2}\left(\max_{k\in\mathcal{K}} \frac{\beta_{k,t}D_k^2}{h_{k,t}^2\left(b_{k,t}-D_kc_k\right)}\right) + \frac{2\left(D-\sum_{k\in\mathcal{K}}\beta_{k,t}D_k\right)^2}{D^2}. \tag{17}$$

Proof: See Appendix A. ■

By repeatedly applying Lemma 1 and collecting terms, we have the optimality gap as follows.

Theorem 1: (**Optimality Gap**) We assume that the total training round is T and the initial global model is $\mathbf{w}_1$. The optimality gap after T training rounds of the OTA FL system can be expressed as follows:

$$\mathbb{E}\left[F\left(\mathbf{w}_{T+1}\right)\right] - F\left(\mathbf{w}^*\right) \leq \Omega_T\left(\boldsymbol{\beta}\right), \tag{18}$$

where

$$\Omega_T\left(\boldsymbol{\beta}\right) = \prod_{t=1}^{T}\psi_t\left(\mathbb{E}\left[F\left(\mathbf{w}_1\right)\right] - F\left(\mathbf{w}^*\right)\right) + \frac{\lambda_1}{l}\left(\sum_{t=1}^{T-1}\prod_{j=t+1}^{T}\psi_j\phi_t + \phi_T\right). \tag{19}$$

B. Problem Transformation

According to Theorem 1, the optimality gap is related to the client selection vector $\boldsymbol{\beta}_t$. If more clients are selected to upload the gradients, the loss decreases faster. However, due to the energy constraint, the induced noise increases when more clients are selected. In contrast from $\boldsymbol{\mathcal{P}}_1$, the energy constraint is added into the problem formulation for $\boldsymbol{\mathcal{P}}_2$. When the total training round is T, we can formulate $\boldsymbol{\mathcal{P}}_2$ as follows:

$$\begin{aligned}\boldsymbol{\mathcal{P}}_2: \quad & \min_{\boldsymbol{\beta}} \quad \Omega_T\left(\boldsymbol{\beta}\right)\\ & \text{s.t.} \quad \beta_{k,t}\in\{0,1\}, \quad \forall k\in\mathcal{K}, 1\leq t\leq T.\end{aligned}$$

In fact, it is difficult to solve $\boldsymbol{\mathcal{P}}_2$ directly due to the stochasticity of the channel state information $\boldsymbol{h}_t$ and the harvested energy $\boldsymbol{e}_t^{\mathrm{arr}}$ for each training round t. We focus on solving the problem via an online pattern based on Lemma 1. The optimality gap $\mathbb{E}[F\left(\mathbf{w}_{t+1}\right)] - F\left(\mathbf{w}^*\right)$ at the t-th training round can be represented by ϕ_t when $\mathbb{E}[F\left(\mathbf{w}_t\right)] - F\left(\mathbf{w}^*\right)$ is known. Therefore, $\boldsymbol{\mathcal{P}}_2$ can be transformed to $\boldsymbol{\mathcal{P}}_3$ to minimize ϕ_t, which can be formulated as follows:

$$\begin{aligned}\boldsymbol{\mathcal{P}}_3: \quad & \min_{\boldsymbol{\beta}_t} \quad \phi_t\left(\boldsymbol{\beta}_t\right)\\ & \text{s.t.} \quad \beta_{k,t}\in\{0,1\}, \quad \forall k\in\mathcal{K}.\end{aligned}$$

Note that $\boldsymbol{\mathcal{P}}_3$ is in essence an NIP. There is an inverse correlation between the optimality gap and the maximum constraint part $\max_{k\in\mathcal{K}}\left(\frac{\beta_{k,t}D_k^2}{h_{k,t}^2(b_{k,t}-D_kc_k)}\right)$. We define the CED coefficient for client k at the t-th training round as follows:

$$q_{k,t} = \frac{\beta_{k,t}D_k^2}{h_{k,t}^2\left(b_{k,t}-D_kc_k\right)}, \tag{20}$$

which is decided by the channel state information $h_{k,t}$, residual battery capacity $b_{k,t}$ and the dataset size D_k. The list of CED coefficients is denoted as $\boldsymbol{Q}_t = [q_{1,t}, q_{2,t}, \ldots, q_{K,t}]$. The maximum value in $\boldsymbol{Q}_t$ is defined as follows: $Q_t^{\max} = \max_{k\in\mathcal{K}} q_{k,t}$. If the selected clients have less available battery, worse channel quality and larger data size, $Q_t^{\max}$ becomes larger.

Algorithm 1 The Joint Channel-Energy-Data (CED)-based Client Selection Algorithm

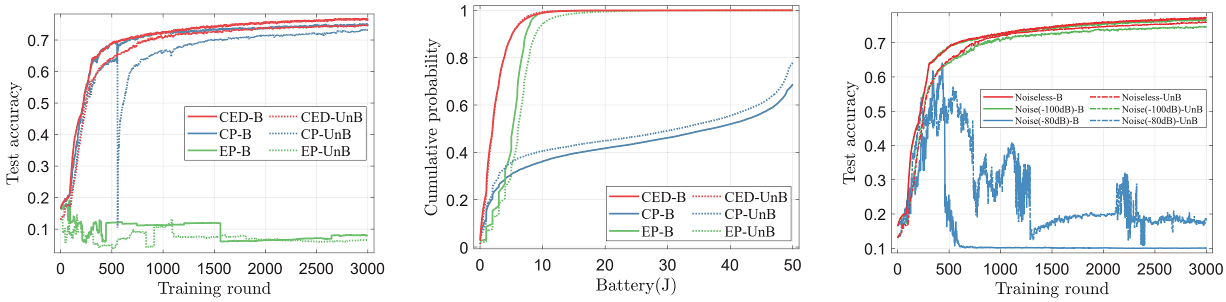
Input: $\boldsymbol{h}_t$, $\boldsymbol{b}_t$, $\boldsymbol{c}_t$, $\boldsymbol{D}$, and D.
Output: $\boldsymbol{\beta}_t$.
for $k = 1, \ldots, K$ **do**
  Calculate $q_{k,t}$ based on (20);
Sort $\boldsymbol{Q}_t$ in ascending order to obtain $\boldsymbol{Q}_t'$;
Let $q_{k,t}'$ be the k-th smallest value in $\boldsymbol{Q}_t'$, and let $\mathcal{S}^{[k]}$ be the client subset decided by the k smallest values of $\boldsymbol{Q}_t'$;
for $k = 1, \ldots, K$ **do**;
  Calculate $\phi_t^{[k]}$ based on (21);
Calculate $k^* = \arg\min_{k\in\mathcal{K}} \phi_t^{[k]}$;
for $k = 1, \ldots, K$ **do**
  if $q_{k,t} \leq q_{k^*,t}$ **then**
    Set $\beta_{k,t} \leftarrow 1$;
  else
    Set $\beta_{k,t} \leftarrow 0$.

IV. Algorithm Design

By analyzing the problem $\boldsymbol{\mathcal{P}}_3$, we can observe that there is a trade-off between $\sum_{k\in\mathcal{K}} D_k\beta_{k,t}$ and $Q_t^{\max}$. If more clients participate in the training, $\sum_{k\in\mathcal{K}} D_k\beta_{k,t}$ will be larger, which will make global loss converge faster. However, if more clients are selected, $Q_t^{\max}$ will be larger, which will lead to a slow convergence. As a result, we need to design an effective client selection scheme.

The proposed solution is referred to Algorithm 1. First, we sort $\boldsymbol{Q}_t$ in ascending order to obtain $\boldsymbol{Q}_t'$. Let $q_{k,t}'$ be the k-th smallest value in $\boldsymbol{Q}_t'$, and let $\mathcal{S}^{[k]}$ be the subset consisting of k clients selected according to the k smallest values of $\boldsymbol{Q}_t'$. There are K possible client selection decisions, and the final client selection decision is based on the following:

$$\phi_t^{[k]} = \frac{s\sigma^2 q_{k,t}'}{2\left(\sum_{i\in\mathcal{S}^{[k]}} D_i\right)^2} + \frac{2\left(D-\sum_{i\in\mathcal{S}^{[k]}} D_i\right)^2}{D^2}. \tag{21}$$



(a) Test accuracy comparison (b) Residual battery comparison (c) The impacts of noise variance

Fig. 1: Performance evaluation among the comparison schemes.

We obtain the indicator k^* by calculating the index of the minimum value of $\phi_t^{[k]}$ as follows:

$$k^* = \arg\min_{k\in\mathcal{K}} \phi_t^{[k]}. \tag{22}$$

The client can be selected if $k \le k^*$ for the sorted queue $\boldsymbol{Q}_t'$, which means $q_{k,t} \le q_{k^*,t}$. The complete process includes the following procedures:

1) At the beginning of the t-th training round, the PS selects a subset of the clients based on Algorithm 1.
2) The PS sends the global model to the selected clients.
3) Local clients update the global model and upload the gradients via OTA computation.
4) The PS obtains the aggregated gradients based on (7).
5) Clients obtain the energy from the energy resources and update the current batter level queue based on formula (10).

The time complexity of Algorithm 1 is mainly determined by the sorting process (see Line 3 therein), which takes $O(K\log K)$ operations. In addition, as Algorithm 1 is conducted T times for convergence, the overall time complexity is given by $\mathcal{O}\left(TK\log K\right)$.

V. Performance Evaluation

In the simulation, there are 40 clients randomly located within the range of a circle with a radius of 250 meters, and the PS is located at $(0, 0, 10)$. The average channel gain for the free-space path loss model is calculated as $\bar{h}_k = G_P G_C \left(\frac{3\times 10^8}{4\pi f_c L_k}\right)^\rho$, where $G_P = 5$ dBi is the antenna gain of the PS, $G_C = 0$ dBi is the antenna gain of clients, $f_c = 915$ MHz is the carrier frequency, L_k is the distance between client k and the PS, and the pass loss exponent is $\rho = 2.7$. The channel gain $h_{k,t}$ during the t-th training round is calculated as $h_{k,t} = \bar{h}_k\gamma_{k,t}$, where $\gamma_{k,t}$ is generated from the Gaussian distribution with zero-mean and unit-variance. The noise variance is set as -100 dB by default. The Poisson parameter $\bar{e}_k$ for client k regarding the arrivals of the harvested energy is set in the range of 0.1 J to 1 J.

We carry out the experiments with the Fashion-MNIST dataset [13]. There are two kinds of settings regarding the data distribution among clients according to [14]. The first setting is a balanced data setting (denoted as 'B') where the number of samples is equal among all clients. The second setting is an unbalanced data setting (denoted as 'UnB'), where the number of samples per client is randomly from [100, 200] for half of the clients and the number of samples per client is randomly from [1000, 2000] for the other clients. We use the four-layer convolutional neural network for training, which is composed of two 5×5 convolution layers, each followed by one 2×2 max pooling layer, one fully connected layer with 50 units, and one softmax layer. The learning rate η is set as 0.01.

We compare the proposed algorithm (denoted as CED) with the following comparison methods:

1) Channel-Prioritized (CP): we select k_{avg} clients with better channels and sufficient energy from $\mathcal{K}$, where k_{avg} is the averaged number of the selected clients of the proposed method.
2) Energy-Prioritized (EP): k_{avg} clients with more energy are selected from $\mathcal{K}$ at the t-th training round.

Fig. 1(a) demonstrates the performance of the test accuracy of the three methods for the two data settings. The proposed CED method is better than the CP method and the EP method for both settings. The reason for this is that the CP method may select clients with less energy, and the EP method may select clients with worse channels, which may introduce more noise and make the gradients deviate from the correct values.

Fig. 1(b) gives the cumulative probability density of the residual battery for 3000 training rounds. The two data settings have similar trends. It can be observed that the residual battery of the CED method is lower than that of the EP method and the CP methods. This is because the communication energy is higher for CED method.

Fig. 1(c) shows the test accuracy for the proposed CED method with different noise settings. It is evident that accuracy decreases when noise increases for both settings. This is because the induced noise makes the gradients deviate from true values for each training round.

VI. Conclusion

The energy management problem is one of the essential challenges in the OTA FL system. Considering the energy harvesting for the OTA FL system, in this paper, we conduct the convergence analysis of the optimality gap with regard to client selection, the energy constraint, and the noise error channel. Based on the convergence analysis results, we formulate the online optimization problem to minimize the optimality gap when jointly considering client selection and energy harvesting. The CED-based method is proposed to optimize client selection. The experiments show that our proposed method surpasses the other benchmarks.

Appendix A
Proof of Lemma 1

As denoted by $\mathbf{o} = \nabla F(\mathbf{w}_t) - \bar{\boldsymbol{r}}_t$, the errors are induced by client selection and noisy channels. When Assumption 1 exists, by incorporating (11) into (12), we have the following:

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) - \eta (\nabla F(\mathbf{w}_t) - \mathbf{o})^T \nabla F(\mathbf{w}_t) + \frac{l\eta^2}{2} \|(\nabla F(\mathbf{w}_t) - \mathbf{o})\|_2^2 . \tag{23}$$

By accessing the expected values of (23) and setting the learning rate to $\eta = \frac{1}{l}$, we see the following:

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t)] - \frac{1}{2l}\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{1}{2l}\mathbb{E}\left[\|\mathbf{o}\|_2^2\right]. \tag{24}$$

Then, according to [4], $\mathbb{E}\left[\|\mathbf{o}\|_2^2\right]$ is bounded as

$$\mathbb{E}\left[\|\mathbf{o}\|_2^2\right] \leq \frac{s\sigma^2}{\left\|\alpha_t \sum_{k\in\mathcal{K}} D_k\beta_{k,t}\right\|_2^2} + \frac{4}{D^2}(D - \sum_{k\in\mathcal{K}}\beta_k D_k)^2(\lambda_1 + \lambda_2\|\nabla F(\mathbf{w}_t)\|_2^2). \tag{25}$$

By incorporating (25) into (24) and subtracting $F(\mathbf{w}^*)$ from both sides of (24), we have the following:

$$\mathbb{E}[F(\mathbf{w}_{t+1})] - F(\mathbf{w}^*) \leq \mathbb{E}[F(\mathbf{w}_t)] - F(\mathbf{w}^*) + \frac{2}{lD^2}(D - \sum_{k\in\mathcal{K}}\beta_k D_k)^2(\lambda_1 + \lambda_2\|\nabla F(\mathbf{w}_t)\|_2^2) + \frac{s\sigma^2}{2l{\alpha_t}^2(\sum_{k=1}^{K} D_k\beta_{k,t})^2} - \frac{1}{2l}\|\nabla F(\mathbf{w}_t)\|_2^2 . \tag{26}$$

According to (8) and (9), the variable α_t at the t-th training round can be expressed as follows:

$$\alpha_t = \min_{k\in\mathcal{K}}\left(\frac{\beta_{k,t}h_{k,t}\sqrt{b_{k,t} - D_k c_k}}{D_k\|\mathbf{g}_{k,t}\|_2}\right). \tag{27}$$

By incorporating (27) into (26) and based on Assumption 2 and Assumption 3, we finally obtain the following:

$$\mathbb{E}[F(\mathbf{w}_{t+1})] - F(\mathbf{w}^*) \leq \psi_t[\mathbb{E}[F(\mathbf{w}_t)] - F(\mathbf{w}^*)] + \frac{\lambda_1}{l}\phi_t. \tag{28}$$

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Numbers JP20K19794 and JP20H00592, in part by the Telecommunications Advancement Foundation (TAF), and in part by the Support Center for Advanced Telecommunications Technology Research, Foundation (SCAT).

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. AISTATS*, Apr. 2017.

[2] P. Li, X. Huang, M. Pan, and R. Yu, “FedGreen: Federated learning with fine-grained gradient compression for green mobile edge computing,” in *Proc. IEEE GLOBECOM*, Dec. 2021.

[3] L. Li, D. Shi, R. Hou, H. Li, M. Pan, and Z. Han, “To talk or to work: Flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices,” in *Proc. IEEE INFOCOM*, May 2021.

[4] 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 Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2020.

[5] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, “Fast convergence algorithm for analog federated learning,” in *Proc. IEEE ICC*, Jun. 2021.

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

[7] X. Fan, Y. Wang, Y. Huo, and Z. Tian, “Joint optimization of communications and federated learning over the air,” *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[8] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, “Over-the-air federated learning from heterogeneous data,” *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, 2021.

[9] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, “Dynamic scheduling for over-the-air federated edge learning with energy constraints,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, 2021.

[10] R. Hamdi, M. Chen, A. B. Said, M. Qaraqe, and H. V. Poor, “Federated learning over energy harvesting wireless networks,” *IEEE Internet Things J.*, vol. 9, no. 1, pp. 92–103, 2021.

[11] B. Güler and A. Yener, “Energy-harvesting distributed machine learning,” in *Proc. IEEE ISIT*, Jul. 2021.

[12] 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, 2018.

[13] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.

[14] H. Liu, X. Yuan, and Y.-J. A. Zhang, “Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7595–7609, 2021.