Update Estimation and Scheduling for Over-the-Air Federated Learning with Energy Harvesting Devices

Furkan Bagci

Furkan Bagci[1], Busra Tegin[1], Mohammad Kazemi[2], and Tolga M. Duman[1]
[1] *Dept. of Electrical and Electronics Engineering, Bilkent University*, Ankara, Turkey
[2] *Dept. of Electrical and Electronic Engineering, Imperial College London*, London, UK
`{bagci, btegin, duman}@ee.bilkent.edu.tr, mohammad.kazemi@imperial.ac.uk`

January 23, 2025

Bilkent University
Department of Electrical and Electronics Engineering

What is Federated Learning (FL)?

Federated Learning

- A machine learning approach where:
  - **Data remains decentralized**
  - Devices collaboratively train a shared global model

Key Components

- **Mobile Users (MUs)**
- **Parameter Server (PS)**

Process

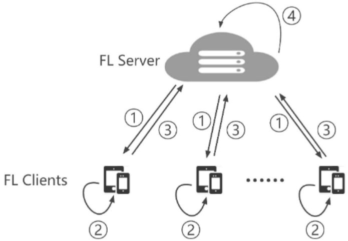


Figure 1: Illustration of a standard FL [1].

1. PS sends the global model to users.
2. Users compute local updates.
3. Updates are aggregated by the PS.
4. Repeat until convergence.

[1] L. Fu, H. Zhang, G. Gao, M. Zhang and X. Liu, "Client Selection in Federated Learning: Principles, Challenges, and Opportunities," in IEEE Internet of Things Journal, vol. 10, no. 24, pp. 21811-21819, 15 Dec.15, 2023.

Why Federated Learning (FL)?

Traditional ML

- Centralized data sharing:
  - **Requires high resources**
  - **Compromises privacy**

Collaborative model training **without sharing local data**

- **Advantages:**
  - Preserves **privacy**
  - Reduces **latency**
  - Improves **learning quality**

Why Over-the-Air (OTA) FL?

Challenge in FL

- Iterative transmission of local updates from mobile users to PS.
- **Key bottleneck:** Communication bandwidth.

Solution: OTA FL

- Leverages **superposition property** of **wireless MAC**
- **Advantages:**
    - Saves bandwidth by avoiding separate transmissions for each user.

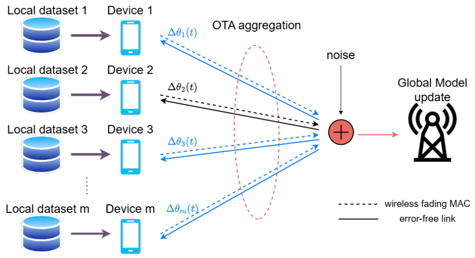


Figure 2: Illustration of OTA FL.

Challenges in OTA FL

Energy Harvesting (EH) Devices

- Uneven and **stochastic** participation in learning.
- Existing studies focus on optimizing energy usage via:
  - Transceiver optimization, receive beamforming design.

Non-i.i.d. Data

- Data heterogeneity impacts:
  - Model convergence [2].
  - Accuracy due to bias in updates.
- Existing works tackle this with:
  - Clustered Sampling [3]
  - Diverse User Selection [4]

- These studies rely on separate transmission of user updates.
- In contrast, our OTA FL setup uses noisy aggregated updates.

[2] X. Li et al., "On the convergence of FedAvg on non-iid data," arXiv preprint, arXiv:1907.02189, 2019. [3] Y. Fraboni et al., "Clustered sampling for client selection in federated learning," ICML, 2021. [4] R. Balakrishnan et al., "Diverse client selection for federated learning via submodular maximization," ICLR, 2022.

Contributions

Diverse User Selection for FL with EH Devices

- **1. Entropy-Based Scheduling:**
  - For known data distributions.
  - Ensures a balanced representation of data labels.
- **2. LSE-Based Scheduling:**
  - For unknown data distributions.
  - Estimates user updates from aggregated signals at the PS.
  - Clusters users based on estimated representations to **enhance diversity** and **eliminate redundant information**.

FL Setup

- M : number of MUs
- K : number of receive antennas
- **Objective:**

$$F(\boldsymbol{\theta}) = \frac{1}{B}\sum_{m=1}^{M}\frac{|B_m|}{B}F_m(\boldsymbol{\theta}), \qquad (1)$$

where:

  - $F_m(\theta)$: Local loss function.
  - $F_m(\boldsymbol{\theta}) = \frac{1}{|B_m|}\sum_{\boldsymbol{u}\in B_m} f(\boldsymbol{\theta}, \boldsymbol{u})$

- Selected users $S(t)$ perform τ iterations of **local SGD:**

$$\boldsymbol{\theta}_m^{i+1}(t) = \boldsymbol{\theta}_m^{i}(t) - \eta_m^{i}(t)\nabla F_m(\boldsymbol{\theta}_m^{i}(t), \xi_m^{i}(t)), \qquad (2)$$

- The m-th user computes the model update as:

$$\Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}_m^{\tau}(t) - \boldsymbol{\theta}_m^{1}(t). \qquad (3)$$

- These updates are transmitted back to the PS for aggregation as:

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{|S(t)|}\sum_{m\in S(t)}\Delta\theta_m(t). \qquad (4)$$

OTA FL Setup

- Using over-the-air transmission over a fading MAC
- The received signal at the k-th antenna of the PS at iteration t is:

$$\boldsymbol{y}_{PS,k}(t) = \sum_{m \in S(t)} \boldsymbol{h}_{m,k}(t) \circ \boldsymbol{x}_m(t) + \boldsymbol{z}_{PS,k}(t), \quad (5)$$

where:

- $\boldsymbol{h}_{m,k}(t)$: i.i.d. channel gain from user m to antenna k with $h^n_{m,k}(t) \sim CN(0, \sigma_h^2)$.
- $\boldsymbol{x}_m(t)$: Signal transmitted by user m.
- $\boldsymbol{z}_{PS,k}(t)$: i.i.d. circularly symmetric AWGN with $z^n_{PS,k}(t) \sim CN(0, \sigma_z^2)$.

- The PS aligns and combines signals from K antennas to mitigate fading effects.

$$\boldsymbol{y}_{PS}(t) = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{m \in S(t)} \boldsymbol{h}_{m,k}(t) \right)^* \circ \boldsymbol{y}_{PS,k}(t), \quad (6)$$

with:

- exact information on the sum of the channel gains

OTA FL Setup

- The n-th symbol of (6) can be partition into three signals[5]

$$y_{PS}^{n}(t) = \underbrace{\sum_{m\in S(t)} \left(\frac{1}{K}\sum_{k=1}^{K} |h_{m,k}^{n}(t)|^2\right)\Delta\theta_{m}^{n,cx}(t)}_{y_{PS}^{n,sig}(t)\,(\text{signal term})}$$
$$+ \underbrace{\frac{1}{K}\sum_{m\in S(t)}\sum_{\substack{m'\in S(t)\\ m'\neq m}}\sum_{k=1}^{K}(h_{m,k}^{n}(t))^{*}\,h_{m',k}^{n}(t)\Delta\theta_{m'}^{n,cx}(t)}_{y_{PS}^{n,int}(t)\,(\text{interference term})}$$
$$+ \underbrace{\frac{1}{K}\sum_{m\in S(t)}\sum_{k=1}^{K}(h_{m,k}^{n}(t))^{*}\,z_{PS,k}^{n}(t).}_{y_{PS}^{n,noise}(t)\,(\text{noise term})} \quad (7)$$

- Recovery of noisy aggregated updates as

$$\Delta\hat{\boldsymbol{\theta}}_{PS}^{n}(t) = \frac{1}{|S(t)|\,\sigma_h^2}\mathrm{Re}\{y_{PS}^{n}(t)\}, \quad (8a)$$

$$\Delta\hat{\boldsymbol{\theta}}_{PS}^{n+N}(t) = \frac{1}{|S(t)|\,\sigma_h^2}\mathrm{Im}\{y_{PS}^{n}(t)\}, \quad (8b)$$

to update the global model, as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t). \quad (9)$$

[5] M. M. Amiri et al., "Blind Federated Edge Learning," IEEE Trans. Wireless Commun., vol. 20, no. 8, pp. 5129-5143, Aug. 2021.

Convergence Analysis

Convergence Rate:

- We have

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right] \leq \left(\prod_{i=0}^{t-1} A(i)\right)\|\boldsymbol{\theta}(0)-\boldsymbol{\theta}^*\|_2^2 + \sum_{j=0}^{t-1} B(j) \prod_{i=j+1}^{t-1} A(i), \tag{10}$$

with

$$A(i) \triangleq 1-\mu\eta(i)\ \left(\tau-\eta(i)(\tau-1)\right),$$

$$B(i) \triangleq \frac{\eta^2(i)\tau^2 G^2}{K} + \frac{\sigma_z^2 N}{\alpha_i^2 K|S(i)|\sigma_h^2}$$
$$+\left(1+\mu(1-\eta(i))\right)\eta^2(i)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} + \eta^2(i)(\tau^2+\tau-1)G^2 + 2\eta(i)(\tau-1)\Gamma$$
$$+\left(\eta^2(t)\tau(\tau-1)LG+\eta(t)\tau\epsilon\right)^2 + \left(\eta^2(t)\tau(\tau-1)LG+\eta(t)\tau\epsilon\right)c,$$

- with ϵ being the gradient approximation error and defined as

$$\epsilon \triangleq \left\| \frac{1}{M}\sum_{m=1}^{M} \nabla F_m(\theta_m(t)) - \frac{1}{|S(t)|}\sum_{m\in S(t)} \nabla F_m(\theta_m(t)) \right\|_2. \tag{11}$$

User Scheduling Strategies

We propose diverse user selection to handle:

- Data heterogeneity
- Stochastic participation

Entropy-Based Scheduling:

- **Goal:** Achieve a uniform representation of data across users.

Methodology:

- Compute the Shannon entropy of label distributions for all available subsets

$$\mathbf{L} = \begin{bmatrix} l_{1,0} & l_{1,1} & \cdots & l_{1,N_c-1} \\ l_{2,0} & l_{2,1} & \cdots & l_{2,N_c-1} \\ \vdots & \vdots & \ddots & \vdots \\ l_{M,0} & l_{M,1} & \cdots & l_{M,N_c-1} \end{bmatrix}.$$

- Select users with the highest combined entropy to ensure diversity

LSE-Based Scheduling

Solution: Least-Squares Estimation (LSE):

- The PS estimates **representative user updates**

Estimation Phase

- Active users transmit their updates without scheduling
- PS stores global updates to create user representations
- Groups users into clusters using **cosine similarities**
- Selects equal users from each cluster for unbiased training

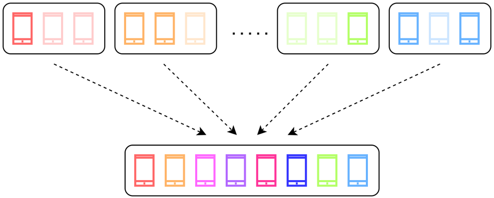


Figure 3: Illustration of clusters and diverse scheduling.

LSE-Based Scheduling

- Define a matrix $\hat{\boldsymbol{\Theta}}_{PS}$, where each row corresponds to the **global model updates**

$$\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j \boldsymbol{\Theta}_j + \boldsymbol{N'}_j$$

$$\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j \begin{bmatrix} \Delta\boldsymbol{\theta}_{j,1} \\ \vdots \\ \Delta\boldsymbol{\theta}_{j,M} \end{bmatrix}_{M\times 2N} + \begin{bmatrix} \boldsymbol{N'}_1 \\ \vdots \\ \boldsymbol{N'}_{2N} \end{bmatrix}^T . \tag{12}$$

where:

  - $\boldsymbol{A}_j$: binary participation vector of size M.
  - $\boldsymbol{\Theta}_j$: matrix with each row representing the local model update from users.
  - $\boldsymbol{N'}_j$: effective noise from MAC fading, AWGN, and combining errors.

- We also define $\boldsymbol{\Theta}_{rep} \in \mathbb{R}^{M\times 2N}$ as

$$\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j(\boldsymbol{\Theta}_{rep} + \boldsymbol{\Theta}_{d,j}) + \boldsymbol{N'}_j, \tag{13}$$

where $\boldsymbol{\Theta}_{d,j} \triangleq \boldsymbol{\Theta}_{rep} - \boldsymbol{\Theta}_j$ and $\boldsymbol{N}^*_j \triangleq \boldsymbol{A}_j\boldsymbol{\Theta}_{d,j} + \boldsymbol{N'}_j$.

- Combining $\hat{\boldsymbol{\Theta}}_{PS,j}$ across T iterations

$$\hat{\boldsymbol{\Theta}}_{PS} = \boldsymbol{A}\boldsymbol{\Theta}_{rep} + \boldsymbol{N}^*, \tag{14}$$

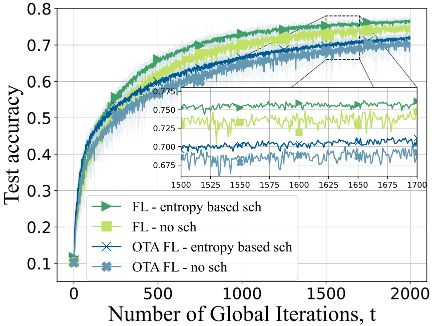
- Solve for $\boldsymbol{\Theta}_{rep}$ using Least-Squares Estimation.
- Using $\boldsymbol{\Theta}_{rep}$, the PS:
  - Captures the **data characteristics** of users.
  - Groups users based on **cosine similarity of representations**.

Numerical Results

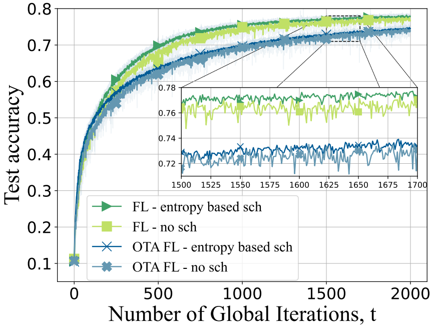
- **MNIST** & **FMNIST**: Single-layer neural network with 2N=7850.
- **CIFAR-10**: Convolutional Neural Network (CNN) with 2N=797,962.
- SGD with a learning rate of 0.05 and a scheduler, $\tau = 5$ and mini-batch size $|\xi_m(t)| = 100$ for MNIST and FMNIST, and $\tau = 3$ and $|\xi_m(t)| = 128$ for CIFAR-10.

- Non-i.i.d. Data Scenarios
  - 1 or 2 classes per user
  - $\boldsymbol{p_m} \sim \mathrm{Dir}_{N_c}(\beta)$ with $\beta \in \{0.1, 0.2\}$
- Wireless Setup
  - $K = 200$
  - Noise variance: $\sigma_h^2 = 1$, and $\sigma_z^2 = 0.1$.

Numerical Results



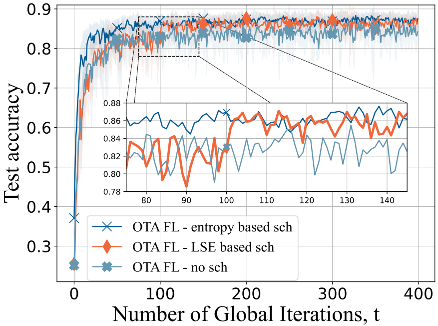
(a) $\beta = 0.1$.



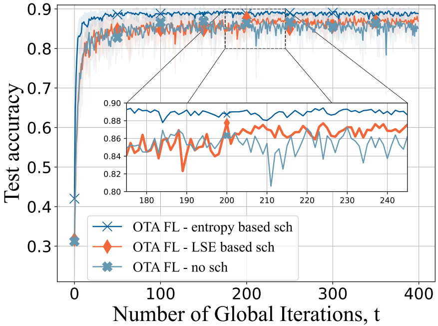
(b) $\beta = 0.2$.

Figure 4: The mean test accuracy of entropy-based scheduling for CIFAR-10 with $M = 100$, $|B_m| = 500$ and $p_e^m(t) = 0.1, \forall m, t$.

Numerical Results



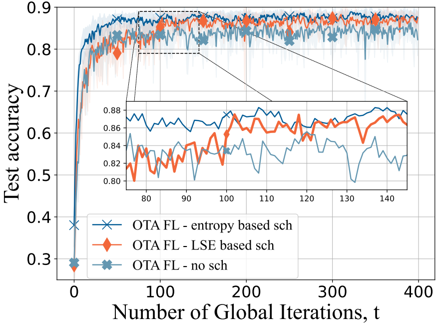
(a) 1 class per user and $T = 100$.



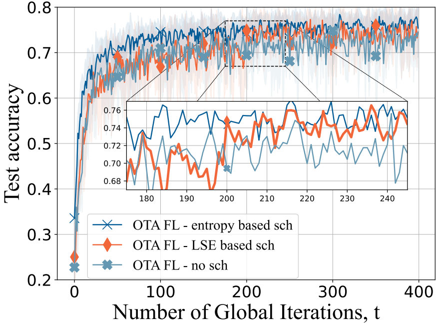
(b) 2 class per user and $T = 200$.

Figure 5: The mean test accuracy for MNIST with $M = 40$, $|B_m| = 1250$ and $p_e^m(t) = 0.25$, $\forall m, t$.

Numerical Results



(a) MNIST, $M = 20$, 1 class per user and $T = 100$.



(b) FMNIST, $M = 40$, 1 class per user and $T = 200$.

Figure 6: The mean test accuracy for MNIST and FMNIST.

Conclusions

- We analyze the convergence rate for the OTA FL with EH devices and demonstrate the effect of user scheduling
- The entropy-based scheduling approach yields higher and more stable accuracy levels.
- User representations can be estimated on the PS side to schedule diverse users, preserve privacy, eliminate redundant update transfers, and improve learning performance.

- **Future Directions**
  - Investigate the effect of estimation strategies under varying scenarios and energy constraints
  - Implement clustered federated learning for the user clusters derived from our estimation.

Update Estimation and Scheduling for Over-the-Air Federated Learning with Energy Harvesting Devices

Furkan Bagci

January 23, 2025

Thank You!

Questions?

Bilkent University
Department of Electrical and Electronics Engineering