Over-the-Air Federated Learning with Energy Harvesting Devices

Ozan Aygün[1], Mohammad Kazemi[1], Deniz Gündüz[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
{ozan, kazemi, duman}@ee.bilkent.edu.tr, d.gunduz@imperial.ac.uk

***Abstract*—We consider federated edge learning among mobile devices that harvest the required energy from their surroundings, and share their updates with the parameter server (PS) through a shared wireless channel. In particular, we consider energy harvesting FL with over-the-air (OTA) aggregation, where the participating devices perform local computations and wireless transmission only when they have the required energy available, and transmit the local updates simultaneously over the same channel bandwidth. In order to prevent bias among the heterogeneous devices, we utilize a weighted averaging with respect to their latest energy arrivals and data cardinalities. We provide a convergence analysis and carry out numerical experiments with different energy arrival profiles, which show that the proposed scheme is robust against heterogeneous energy arrivals in error-free scenarios while having less than 10% performance loss for fading channels.**

***Index Terms*—Federated learning, energy harvesting devices, wireless communications, machine learning.**

I. Introduction

Developments in Internet-of-things (IoT) paradigm have helped machine learning (ML) approaches to be used in many domains such as healthcare, automation, and forecasting, thanks to the endless data collection capabilities. While mobile devices are at the center of attention for collecting data, traditional ML approaches require the collected data to be assembled in a cloud server for model training. However, this approach may not be feasible due to several reasons. Firstly, the participants are typically reluctant to share their private data; secondly, sending all the data to a server has a high communication cost, particularly in bandwidth and energy-limited scenarios. Finally, latency can be a critical limitation for time-sensitive applications [1]. *Federated learning* (FL) is a recently emerging framework that aims to mitigate these issues, where the participating devices perform model training with local data and send their parameter updates to the parameter server (PS), which orchestrates the learning process, instead of sharing the local data itself to preserve privacy [2], [3].

In FL, participating devices called mobile devices (MDs) can be selected based on their available energy, computing capability, and quality of their channel to the PS [4]. Before the local training, the PS sends the global model to the MDs. Selected MDs perform stochastic gradient descent (SGD) iterations using their local dataset. After completion, a subset of the MDs shares their model updates with the PS, where the model aggregation is performed to obtain the updated global model. These steps are repeated either for a prescribed number of iterations or until a certain condition is met. Recent studies on FL include investigating the effects of data heterogeneity [2], [5], [6], design of communication-efficient approaches [7]–[12], and latency and power analysis [13], [14].

Even though FL has many potential benefits in terms of privacy and communication cost, bandwidth limitations and adverse channel conditions in wireless setups may threaten its feasibility in certain practical scenarios. To reduce the required bandwidth in FL, over-the-air (OTA) aggregation has become the *de facto* approach where the same bandwidth is shared by all the participating MDs, enabling the aggregation of gradients during the transmission [9]. The adverse channel effects can be alleviated using multiple receive antennas and combining techniques at the PS [15]–[18].

Despite the success of FL in practical scenarios, the energy consumption and carbon footprint of MDs for training and sharing their local models create serious concerns about the sustainability of future smart systems [19]. As a more sustainable approach, energy harvesting devices, which can acquire energy from their surroundings [20], have been widely considered for mobile networks. These devices are typically equipped with a rechargeable battery to store the harvested energy and perform the required computations and communications if they have available energy in their battery.

Energy harvesting communication devices have been previously studied in detail from different perspectives. The results include optimal transmission policies [21]–[23], and channel capacity computation for unit-sized battery [24]. FL with energy harvesting MDs has also been considered lately [25], [26]. However, existing approaches do not consider the wireless channel effects or OTA aggregation in energy harvesting FL setups, which constitutes the basis of this work.

To examine the performance of FL with energy harvesting in practical settings, we introduce OTA FL with energy harvesting MDs. In this setting, the participating MDs perform local SGD iterations and transmit their gradients using wireless

The authors acknowledge support from TUBITAK through CHIST-ERA project SONATA (CHIST-ERA-20-SICT-004, funded by TUBITAK, Turkey Grant 221N366 and EPSRC, UK Grant EP/W035960/1).

Ozan Aygün's research is also supported by Turkcell A.S. within the framework of 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority of Turkey.

links simultaneously over the same frequency band. Using OTA aggregation and combining techniques, the PS updates the global model based on the received signal, and the updated model is sent back to the users for the next global iteration. We compare the performance of our setup with the error-free scenarios and conventional FL using different energy arrival profiles. Numerical and experimental results show that even under energy harvesting limitations, the proposed algorithm can perform well for practical channel models with large number of users with convergence guarantees.

The rest of the paper is organized as follows. In Section II, we introduce the FL setup as well as the energy harvesting processes at the devices with different energy arrival profiles. In Section III, we study the OTA communication model of FL with MDs that have intermittent energy arrivals. In Section IV, a convergence analysis of energy harvesting FL is presented under certain convexity assumptions on the loss function. We give several numerical results in Section V, and conclude the paper in Section VI.

II. System Model

A. FL Setup

The main goal in FL is to minimize a global loss function $F(\boldsymbol{\theta})$ with respect to the model weights $\boldsymbol{\theta} \in \mathbb{R}^{2N}$, where $2N$ is the dimension of the weights in the model. Our system has M single-antenna MDs and a PS equipped with K antennas. Each MD has a dataset $\mathcal{B}_m$ with cardinality $|\mathcal{B}_m|$, and we define $B \triangleq \sum_{m=1}^{M} |\mathcal{B}_m|$ as the number of total data samples. We define the global loss function as

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

where $F_m(\boldsymbol{\theta}) \triangleq \frac{1}{|\mathcal{B}_m|} \sum_{u \in \mathcal{B}_m} f(\boldsymbol{\theta}, u)$, with $f(\boldsymbol{\theta}, u)$ corresponding to the loss of the u-th data sample.

In every global iteration, the MDs perform τ local SGD iterations using their local data to obtain model updates that need to be shared with the PS for the global aggregation. The SGD steps at the m-th MD at the i-th local and t-th global iteration are performed as

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

where $\eta_m^i(t)$ is the learning rate, $\nabla F_m(\boldsymbol{\theta}_m^i(t), \boldsymbol{\xi}_m^i(t))$ is the unbiased local gradient estimate for the local weights $\boldsymbol{\theta}_m^i(t)$ with the randomly sampled batch $\boldsymbol{\xi}_m^i(t)$ from the dataset $\mathcal{B}_m$, i.e., $\mathbb{E}_\xi \left[\nabla F_m(\boldsymbol{\theta}_m(t), \boldsymbol{\xi}_m(t))\right] = \nabla F_m(\boldsymbol{\theta}_m(t))$, where the expectation is over the random batch of data samples.

Having computed the local SGD steps, MDs calculate their model difference to be shared with the PS as

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

In the case where all the devices participate in the global aggregation with error-free transmission, the PS computes

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \sum_{m=1}^{M} p_m(t) \Delta \boldsymbol{\theta}_m(t), \tag{4}$$

where $\boldsymbol{\theta}_{PS}(t)$ represents the model weight vector at the PS at the t-th global iteration and $p_m(t) = \frac{|\mathcal{B}_m|}{\sum_{m=1}^{M} |\mathcal{B}_m|}$ denotes the ratio of the number of data samples of the m-th device to the total number of samples participating in the aggregation. Note that the denominator can change depending on the number of participating devices. The updated global weights at the PS are shared with the MDs for the next global iteration.

Global aggregation can also be performed via OTA aggregation, where the local model updates can be transmitted over a shared wireless medium to the PS, whose output for the k-th receive antenna becomes [1]

$$\boldsymbol{y}_{PS,k}(t) = \sum_{m=1}^{M} \boldsymbol{h}_{m,k}(t) \circ \boldsymbol{x}_m(t) + \boldsymbol{z}_{PS,k}(t), \tag{5}$$

where $\boldsymbol{x}_m(t)$ is the transmitted signal from the m-th MD, $\circ$ denotes the element-wise product, $\boldsymbol{z}_{PS,k}(t) \in \mathbb{C}^N$ is the circularly symmetric additive white Gaussian noise (AWGN) vector with independent and identically distributed (i.i.d.) entries with zero mean and variance of σ_z^2; i.e., $z_{PS,k}^n(t) \sim \mathcal{CN}(0, \sigma_z^2)$. The channel coefficients are given as $\boldsymbol{h}_{m,k}(t) = \sqrt{\beta_m}\ \boldsymbol{g}_{m,k}(t)$, where $\boldsymbol{g}_{m,k}(t) \in \mathbb{C}^N$ with each entry $g_{m,k}^n(t) \sim \mathcal{CN}(0, \sigma_h^2)$ (i.e., Rayleigh fading), β_m is the large-scale fading coefficient modeled as $\beta_m = (d_m)^{-p}$, where p denotes the path loss exponent, and d_m represents the distance between the m-th MD and the PS.

B. Energy Harvesting Devices

We consider OTA FL with energy harvesting devices where each MD has a unit battery. The MDs harvest either unit energy, or no energy at all from various sources such as solar, kinetic, or RF energy in every global iteration. For simplicity, we assume that τ local SGD steps and the transmission of gradients to the PS cost a unit amount of energy.

We denote the binary energy arrival process of the m-th MD at the t-th global iteration as $E_m(t)$. If $E_m(t) = 1$, then the m-th MD receives enough energy to participate in the global iteration t. $E_m(t) = 0$, if no energy is harvested. We also define the elapsed time between the current iteration and the previous energy arrival as $\lambda_m(t) = \max_{t':t'<t, E_m(t')=1} t'$. Lastly, for a given t, we define a quantity called the cooldown multiplier as $c_m(t) = t - \lambda_m(t)$, which represents the number of iterations for which the m-th MD has not been harvesting energy.

We investigate the use of MDs with stochastic energy arrival profiles, where the harvested energy has an underlying probability distribution, and the MDs have no prior information about the next energy arrival time. Note that the MDs do not know the underlying distribution of the stochastic process. We will consider two different stochastic energy arrival processes: Bernoulli and uniform energy arrivals.

[1] Note that OTA aggregation can be implemented using orthogonal frequency division multiplexing (OFDM) in practice.

1) Bernoulli: At the t-th global iteration, the m-th MD receives energy with probability $\alpha_m(t)$, i.e.,

$$E_m(t) = \begin{cases} 1 & \text{with probability} \quad \alpha_m(t), \\ 0 & \text{with probability} \quad 1-\alpha_m(t). \end{cases} \tag{6}$$

2) Uniform: Global iterations are divided into blocks of length T_m, and the m-th MD receives energy once for every T_m iterations. This means that with probability 1, an energy arrival is observed in $\{t, \ldots, t+T_m-1\}$.

III. OTA FL with Energy Harvesting

We now describe the proposed FL scheme with energy harvesting MDs where the gradients are sent through wireless channels using OTA aggregation. Since the mobile devices do not always have sufficient energy to perform local SGD computations or gradient transmissions, only the MDs that have harvested enough energy, i.e., those with $E_m(t)=1$ can participate in the t-th global iteration. We define $\mathcal{S}(t)$ as the set of devices participating in the t-th global iteration.

Before each training round, the MDs receive the current global model $\boldsymbol{\theta}_{PS}(t)$ from the PS. If an MD has sufficient energy to participate in the t-th iteration, the SGD calculations are performed. Then, based on the cooldown multiplier of each MD, the weighted model differences are computed using

$$\Delta\boldsymbol{\theta}_m^s(t) = C_m(t)\Delta\boldsymbol{\theta}_m(t), \tag{7}$$

where $C_m(t) = p_m(t)c_m(t)$, and $\Delta\boldsymbol{\theta}_m^s(t)$ denotes the scaled model differences for the m-th MD at the t-th global iteration. Considering error-free transmission of the scaled gradients, the PS performs global update for the next iteration as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t), \tag{8}$$

where $\Delta\boldsymbol{\theta}_{PS}(t)$ is defined as

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{C(t)} \sum_{m\in\mathcal{S}_t} \Delta\boldsymbol{\theta}_m^s(t) \tag{9}$$

with $C(t) = \sum_{m\in\mathcal{S}_t} C_m(t)$, which is assumed to be known by the PS [25]. The reader is referred to [27] and the references therein for related algorithms to estimate the number of participating users.

We now consider the OTA aggregation of the local model differences. The PS receives a noisy target signal due to the wireless channel and the noise. In the proposed scheme, we assume perfect channel state information (CSI) at the receiver side and no CSI at the MDs.

For a more spectrally efficient approach, the model differences are written in terms of a complex signal $\Delta\boldsymbol{\theta}_m^{s,cx}(t) \in \mathbb{C}^N$ by grouping the symbols into its real and imaginary parts as

$$\Delta\boldsymbol{\theta}_m^{s,re}(t) \triangleq \left[\Delta\theta_m^{s,1}(t), \Delta\theta_m^{s,2}(t), \ldots, \Delta\theta_m^{s,N}(t)\right]^T, \tag{10a}$$

$$\Delta\boldsymbol{\theta}_m^{s,im}(t) \triangleq \left[\Delta\theta_m^{s,N+1}(t), \Delta\theta_m^{s,N+2}(t), \ldots, \Delta\theta_m^{s,2N}(t)\right]^T. \tag{10b}$$

For the k-th antenna, the PS receives the following signal

$$\boldsymbol{y}_{PS,k}(t) = \sum_{m\in\mathcal{S}_t} \boldsymbol{h}_{m,k}(t) \circ \Delta\boldsymbol{\theta}_m^{s,cx}(t) + \boldsymbol{z}_{PS,k}(t), \tag{11}$$

With the assumption that perfect CSI is available at the receiver side, combining can be performed as (see [15])

$$\boldsymbol{y}_{PS}(t) = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m\in\mathcal{S}_t} \boldsymbol{h}_{m,k}(t)\Big)^* \circ \boldsymbol{y}_{PS,k}(t). \tag{12}$$

For the n-th symbol, the combined signal becomes

$$\begin{aligned} y_{PS}^n(t) = &\underbrace{\sum_{m\in\mathcal{S}_t}\Big(\frac{1}{K}\sum_{k=1}^{K}|h_{m,k}^n(t)|^2\Big)\Delta\theta_{m,s}^{n,cx}(t)}_{y_{PS}^{n,sig}(t) \text{ (signal term)}} \\ &+ \underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t \\ m'\neq m}}\sum_{k=1}^{K}(h_{m,k}^n(t))^* h_{m',k}^n(t)\Delta\theta_{m',s}^{n,cx}(t)}_{y_{PS}^{n,int}(t) \text{ (interference term)}} \\ &+ \underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}(h_{m,k}^n(t))^* z_{PS,k}^n(t)}_{y_{PS}^{n,noise}(t) \text{ (noise term)}}. \end{aligned} \tag{13}$$

We recover the aggregated model differences from the received signal as

$$\Delta\hat{\theta}_{PS}^n(t) = \frac{1}{C(t)\sigma_h^2\beta}\,\mathrm{Re}\{y_{PS}^n(t)\}, \tag{14a}$$

$$\Delta\hat{\theta}_{PS}^{n+N}(t) = \frac{1}{C(t)\sigma_h^2\beta}\,\mathrm{Im}\{y_{PS}^n(t)\}. \tag{14b}$$

Finally, the global update can be performed as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{15}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \left[\Delta\hat{\theta}_{PS}^1(t)\ \Delta\hat{\theta}_{PS}^2(t)\ \cdots\ \Delta\hat{\theta}_{PS}^{2N}(t)\right]^T$.

IV. Convergence Analysis

In this section, we analyze the convergence rate of the proposed algorithm by upper bounding the difference between the global loss of the FL model and the optimal model.

We denote the minimum local loss as F_m^*, the optimal weights of the model as $\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, and the minimum total loss function is given as $F^* = F(\boldsymbol{\theta}^*)$. The dataset bias is defined as $\Gamma \triangleq F^* - \sum_{m=1}^{M} p_m F_m^* \geq 0$. Moreover, it is assumed that the same learning rate is used accross different MDs, i.e., $\eta_m^i(t) = \eta(t)$.

Assumption 1. *Squared l_2 norm of the local stochastic gradients are bounded; i.e.,*

$$\mathbb{E}_\xi\Big[\,\|\nabla F_m(\boldsymbol{\theta}_m(t), \boldsymbol{\xi}_m(t))\|_2^2\Big] \leq G^2, \tag{16}$$

which implies $\forall n \in [2N]$, $\mathbb{E}_\xi[\nabla F_m(\theta_m^n, \xi_m^n(t))] \leq G$.

Assumption 2. *Local loss functions are assumed to be L-smooth and μ-strongly convex; i.e., $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{2N}$, $\forall m \in [M]$,*

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \leq \langle \boldsymbol{a}-\boldsymbol{b}, \nabla F_m(\boldsymbol{b})\rangle + \frac{L}{2}\|\boldsymbol{a}-\boldsymbol{b}\|_2^2, \tag{17}$$

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \geq \langle \boldsymbol{a}-\boldsymbol{b}, \nabla F_m(\boldsymbol{b})\rangle + \frac{\mu}{2}\|\boldsymbol{a}-\boldsymbol{b}\|_2^2. \tag{18}$$

Theorem 1. *In energy harvesting OTA FL with Bernoulli energy arrivals $\alpha_m = \alpha$ and equal data distribution $p_m = p, \forall m \in [M]$, for $0 \leq \eta(t) \leq min\{1, \frac{1}{\tau\mu}\}$, we can upper bound the model difference between the global and the optimal weights as*

$$\mathbb{E}\left[\, \|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2 \,\right] \leq \left(\prod_{a=1}^{t-1} X(a)\right) \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=1}^{t-1} Y(b) \prod_{a=b+1}^{t-1} X(a), \quad (19)$$

where $X(a) = (1 - \mu\eta(a)\,(\tau - \eta(a)(\tau-1)))$ and

$$\begin{aligned} Y(a) =\ & \tau^2 G^2 \eta^2(a) \sum_{m_1 \in \mathcal{S}_a} \sum_{m_2 \in \mathcal{S}_a} A(m_1, m_2) \\ & + \frac{\tau^2 G^2 \eta^2(a)}{K\bar{\beta}^2} \sum_{m \in \mathcal{S}_a} \sum_{\substack{m' \in \mathcal{S}_a \\ m' \neq m}} \beta_m \beta_{m'} + \frac{\sigma_z^2 N}{p^2 K \sigma_h^2} \sum_{m \in \mathcal{S}_a} \frac{\beta_m}{\bar{\beta}^2} \\ & + (1 + \mu(1 - \eta(a)))\, \eta^2(a) G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ & + \eta^2(a)(\tau^2 + \tau - 1)G^2 + 2\eta(a)(\tau - 1)\Gamma. \qquad (20) \end{aligned}$$

with $A(m_1, m_2) = \left(1 - \frac{\beta_{m_1}}{\bar{\beta}} - \frac{\beta_{m_2}}{\bar{\beta}} + \frac{(M\alpha+1)(K+1)\beta_{m_1}\beta_{m_2}}{M\alpha K\bar{\beta}^2}\right)$.

Proof: Define an auxiliary variable $\boldsymbol{v}(t+1) \triangleq \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t)$, where $\Delta\boldsymbol{\theta}_{PS}(t)$ is defined in (9). Then, we have

$$\begin{aligned} \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{\theta}^*\|_2^2 &= \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &= \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\|_2^2 + \|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad + 2\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^* \rangle. \qquad (21) \end{aligned}$$

In the following lemmas, we provide upper bounds for (21).

Lemma 1. $\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\big\|_2^2\Big]$

$$\begin{aligned} &\leq \tau^2 G^2 \eta^2(t) \sum_{m_1 \in \mathcal{S}_t} \sum_{m_2 \in \mathcal{S}_t} A(m_1, m_2) + \frac{\sigma_z^2 N}{p^2 K \sigma_h^2} \sum_{m \in \mathcal{S}_t} \frac{\beta_m}{\bar{\beta}^2} \\ &\quad + \frac{\tau^2 G^2 \eta^2(t)}{K\bar{\beta}^2} \sum_{m \in \mathcal{S}_t} \sum_{\substack{m' \in \mathcal{S}_t \\ m' \neq m}} \beta_m \beta_{m'}. \qquad (22) \end{aligned}$$

Proof: See Appendix A. ■

Lemma 2. $\mathbb{E}\Big[\big\|v(t+1) - \boldsymbol{\theta}^*\big\|_2^2\Big]$

$$\begin{aligned} &\leq (1 - \mu\eta(t)(\tau - \eta(t)(\tau-1)))\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\big\|_2^2\Big] \\ &\quad + (1 + \mu(1 - \eta(t)))\, \eta^2(t) G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ &\quad + \eta^2(t)(\tau^2 + \tau - 1)G^2 + 2\eta(t)(\tau-1)\Gamma. \qquad (23) \end{aligned}$$

Proof: The proof follows the same line as in Lemma 2 in [15]. ■

Lemma 3. $\mathbb{E}\left[\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^* \rangle\right] = 0.$

Proof: The derivation is the same as in Lemma 3 in [18] by using the independence between local updates and individual channel realizations. ■

The theorem is concluded after applying recursion to the results of Lemmas 1-3. ■

Corollary 1. *Using Assumption 2, the global loss can be upper bounded after T global iterations as*

$$\begin{aligned} \mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T)) - F^*\right] &\leq \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T) - \boldsymbol{\theta}^*\|_2^2\right] \\ &\leq \frac{L}{2}\left(\prod_{n=1}^{T-1} X(n)\right) \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \frac{L}{2}\sum_{p=1}^{T-1} Y(p) \prod_{n=p+1}^{T-1} X(n). \qquad (24) \end{aligned}$$

Assuming $\tau = 1, \beta_m = 1, \forall m \in [M], \eta(t) = \eta, \forall t$ and knowing that $K \gg M$, we get

$$\begin{aligned} \mathbb{E}\left[F\big(\boldsymbol{\theta}_{PS}(T)\big)\right] - F^* &\approx \frac{L}{2}\big(1 - \mu\eta\big)^T \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad + \frac{L}{2\mu\eta}\Big(2\eta^2 G^2 + \frac{\sigma_z^2 N}{p^2 K \sigma_h^2}\Big)\Big(1 - \big(1 - \mu\eta\big)^T\Big). \qquad (25) \end{aligned}$$

Remark. The noise term in $Y(t)$ does not depend on $\eta(t)$, so we have $\lim_{t\to\infty} \mathbb{E}[F(\boldsymbol{\theta}_{PS}(t))] - F^* \neq 0$ even though $\lim_{t\to\infty} \eta(t) = 0$. As expected, having more receive antennas and more data contribution from devices increases the convergence rate, whereas the model size and the noise variance have negative effects.

V. Simulation Results

We consider an FL environment with $M = 40$ MDs and a PS with $K = 5M$ receive antennas. MDs are spread around the PS randomly in such a way that their distances to the PS is uniformly distributed between 0.5 and 2 units.

We use the CIFAR-10 dataset [28] with Adam optimizer [29], and consider an i.i.d. data distribution where the data samples are randomly and equally distributed among MDs. The architecture presented in [15] is used with $2N = 307498$.

We study the performance of conventional FL (without any communication constraints), OTA FL where all the MDs have available energy to participate at all iterations, and energy harvesting FL where MDs have intermittent energy arrivals with both error-free and OTA aggregation schemes. To make a comparison with the previous studies, we also consider the setup used in [25] with Bernoulli energy arrivals, which corresponds to the energy harvesting FL setup with no channel errors without any normalization at the PS with respect to the cooldown multipliers. Moreover, the MDs are divided into 4 equal-sized groups with different energy profiles. For Bernoulli energy arrivals, we have $\alpha_m(t) \in \{1, 1/5, 1/10, 1/20\}$, and for uniform energy arrivals, we have $T_m \in \{1, 5, 10, 20\}$ for MDs in 4 groups as in [25]. The training is performed for $T = 1000$ global iterations for $\tau = 1$, and $T = 400$ for $\tau = 3$ with mini-batch size $|\boldsymbol{\xi}_m(t)| = 128$, the path loss exponent $p = 4$, $\sigma_h^2 = 1$, and $\sigma_z^2 = 1$.

Accuracy plots for the case with Bernoulli energy arrival profiles with $\tau = 1$ and $\tau = 3$ are presented in Figs. 1 and 2, respectively. The results show that the energy harvesting FL with error-free links has a convergence rate close to that of FL with full participation, and that adding a normalization term with respect to the cooldown multipliers leads to a faster

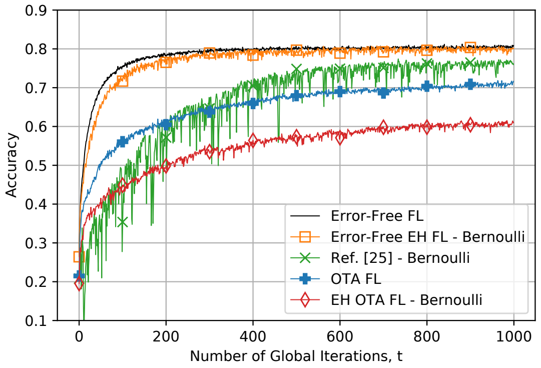

Fig. 1: Test accuracy for $\tau = 1$

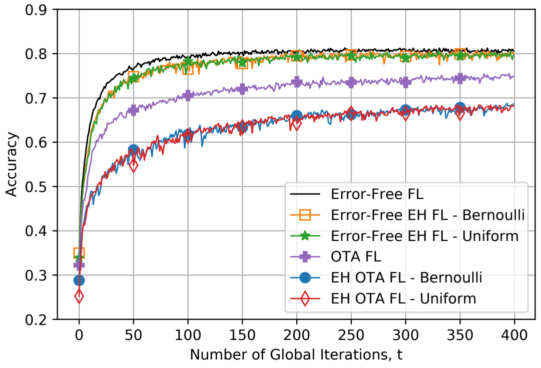

Fig. 2: Test accuracy for $\tau = 3$

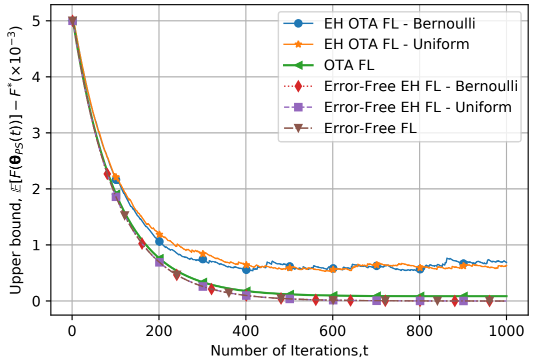

Fig. 3: Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(t)) - F^*\right]$

convergence and less fluctuations compared to the results in [25]. Moreover, the performance of OTA FL is very similar to the scenario used in [25] with error-free links. It can be seen that even though the links are wireless, the gap in the performance can be compensated as the number of global iterations increases. One reason is that the increased number of receive antennas at the PS can reduce the adverse affects of the small-scale fading and noise. Increasing τ achieves a better performance with faster convergence at the cost of making more computations at the edge. It can also be observed that performances of Bernoulli and uniform arrivals are very close to each other due to the energy arrival profile similarities.

In Fig. 3, we numerically evaluate the convergence rates of the scenarios in Fig. 2, using the expression in (24) with $M = 40, 2N = 307498, L = 10, \mu = 1, \tau = 1, G^2 = 1, \eta(t) = 10^{-2} - 10^{-6}t, \sigma_z^2 = 5, \sigma_h^2 = 1, K = M, \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. We observe a close convergence rate between the conventional FL and the error-free energy harvesting FL as expected due to weighted averaging operation with respect to the cooldown multipliers. Energy harvesting FL with OTA aggregation has a slower convergence rate when compared to the others because of the wireless channel effects as well as the decreased number of participants at each iteration due to energy harvesting devices. Since the number of participating devices $|\mathcal{S}_t|$ is random in nature, and it affects $C(t)$, the shifts and fluctuations on the convergence rates are observed.

VI. Conclusions

We study OTA FL with energy harvesting devices with intermittent and heterogeneous energy arrivals. Our framework consists of local SGD computations at the MDs that have available energy, and OTA aggregation of the gradients over a shared wireless medium. A comparison of the performance of the OTA FL with energy harvesting devices through neural network simulations and an analysis of its convergence rate through numerical experiments are performed. The results with different energy profiles demonstrate that performing a weighted averaging using the latest energy arrival and dataset cardinality in energy harvesting FL can give a similar performance to the full-participation scheme in both error-free and OTA cases. As a future direction, one can investigate set-ups with different battery capacities, and optimize the amount of power to allocate for computation versus transmission.

Appendix A

We can write $\Delta\hat{\theta}_{PS}^n(t) = \sum_{p=1}^3 \Delta\hat{\theta}_{PS,p}^n(t)$, for the n-th symbol using (13), because of the i.i.d. of channel realizations, we obtain

$$\mathbb{E}\big[||\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)||_2^2\big] = \mathbb{E}\big[\big\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t) - \Delta\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]$$
$$= \sum_{n=1}^{2N} (\mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,1}^n(t) - \Delta\theta_{PS}^n(t)\big)^2\big] + \sum_{p=2}^{3} \mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,l}^n(t)\big)^2\big]. \quad (26)$$

Lemma 4. $\sum_{n=1}^{2N} \mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,1}^n(t) - \Delta\theta_{PS}^n(t)\big)^2\big]$

$$\leq \sum_{n=1}^{2N} \sum_{m_1 \in \mathcal{S}_t} \sum_{m_2 \in \mathcal{S}_t} A(m_1, m_2) \mathbb{E}\big[\Delta\theta_{m_1}^n(t) \Delta\theta_{m_2}^n(t)\big]. \quad (27)$$

where $A(m_1, m_2) = \Big(1 - \frac{\beta_{m_1}}{\bar{\beta}} - \frac{\beta_{m_2}}{\bar{\beta}} + \frac{2 + (M\alpha - 1)(K-1)\beta_{m_1}\beta_{m_2}}{M\alpha K\bar{\beta}^2}\Big)$.

Proof: For a single symbol, we can write

$$\mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,1}^n(t) - \Delta\theta_{PS}^n(t)\big)^2\big]$$
$$= \mathbb{E}\Big[\frac{1}{C(t)^2} \sum_{m_1 \in \mathcal{S}_t} \sum_{m_2 \in \mathcal{S}_t} C_{m_1}(t) C_{m_2}(t) \Delta\theta_{m_1}^n(t) \Delta\theta_{m_2}^n(t)$$
$$\times \Big(1 - \frac{1}{K\sigma_h^2\bar{\beta}} \sum_{k_1=1}^{K} |h_{m_1,k_1}^n(t)|^2 - \frac{1}{K\sigma_h^2\bar{\beta}} \sum_{k_2=1}^{K} |h_{m_2,k_2}^n(t)|^2$$
$$+ \frac{1}{K^2\sigma_h^4\bar{\beta}^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} |h_{m_1,k_1}^n(t)|^2 |h_{m_2,k_2}^n(t)|^2\Big)\Big]. \quad (28)$$

Using $C_m(t) \leq p$ and $C^2(t) \leq p^2$ and utilizing the i.i.d. channel realizations result in (27). ■

Lemma 5. $\sum_{n=1}^{2N} \mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] \leq \sum_{m\in\mathcal{S}_t} \sum_{\substack{m'\in\mathcal{S}_t \\ m'\neq m}} \frac{\beta_m\beta_{m'}}{K\bar{\beta}^2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_{m'}(t)\|_2^2\big].$

Proof: For the real part, using the independence of channels for different m's and k's, we obtain

$$\begin{aligned}
\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t} \sum_{\substack{m'\in\mathcal{S}_t \\ m'\neq m}} \frac{1}{C(t)K\sigma_h^2\bar{\beta}} \\
&\quad \times \sum_{k=1}^{K} \mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* h^n_{m',k}(t) C_{m'}(t)\Delta\theta^{n,cx}_{m'}(t)\right\}\Big)^2\Big] \\
&\leq \mathbb{E}\Big[\sum_{m\in\mathcal{S}_t} \sum_{\substack{m'\in\mathcal{S}_t \\ m'\neq m}} \frac{\beta_m\beta_{m'}}{2K\bar{\beta}^2}\big(\big(\Delta\theta^n_{m'}(t)\big)^2 + \big(\Delta\theta^{n+N}_{m'}(t)\big)^2 \\
&\quad + \Delta\theta^n_m(t)\Delta\theta^n_{m'}(t) - \Delta\theta^{n+N}_m(t)\Delta\theta^{n+N}_{m'}(t)\big)\Big] \qquad (29)
\end{aligned}$$

We obtain a similar expression for $N+1 \leq n \leq 2N$, and summing the two parts concludes the lemma. ■

Lemma 6. $\sum_{n=1}^{2N} \mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] \leq \frac{\sigma_z^2 N}{p^2 K\sigma_h^2} \sum_{m\in\mathcal{S}_t} \frac{\beta_m}{\bar{\beta}^2}.$

Proof: The first half of the signal yields to

$$\begin{aligned}
&\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] \\
&= \mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K} \frac{1}{C(t)K\sigma_h^2\bar{\beta}} \mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^n_{PS,k}(t)\right\}\Big)^2\Big] \\
&\leq \frac{1}{p^2K^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}\big(\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^{i,n}_{PS,k}(t)\right\}\big)^2\Big] \\
&\overset{(a)}{=} \frac{\sigma_z^2}{2p^2K\sigma_h^2}\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{\beta}^2}. \qquad (30)
\end{aligned}$$

where (a) is obtained using the independence between the channel realizations and the noise. The result also holds for $N+1 \leq n \leq 2N$. Summing with respect to all the symbols completes the proof. ■

The proof is completed using Assumption 1 and (2), and summing up the results in Lemmas 4-6.

References

[1] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.

[2] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. 20th Intl. Conf. on Artif. Intell. Stats. (AISTATS)*, pp. 1273–1282, 2017.

[4] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to earn at the edge," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 14–19, 2020.

[5] W. Zhang, X. Wang, P. Zhou, W. Wu, and X. Zhang, "Client selection for federated learning with non-iid data in mobile edge computing," *IEEE Access*, vol. 9, pp. 24 462–24 474, 2021.

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

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

[8] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2021.

[9] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[10] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Academy of Sciences*, vol. 118, no. 17, 2021.

[11] B. Tegin and T. M. Duman, "Blind federated learning at the wireless edge with low-resolution ADC and DAC," *IEEE Trans. on Wireless Commun.*, 2021.

[12] ——, "Federated learning over time-varying channels," Madrid, Spain, Dec. 2021.

[13] C. T. Dinh *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, 2020.

[14] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, 2021.

[15] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, 2021.

[16] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, pp. 1–16, 2021.

[17] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[18] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Hierarchical over-the-air federated edge learning," in *2022 IEEE Intl. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022.

[19] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[20] S. Ulukus *et al.*, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, 2015.

[21] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, 2011.

[22] B. Gurakan, O. Ozel, J. Yang, and S. Ulukus, "Energy cooperation in energy harvesting communications," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4884–4898, 2013.

[23] O. Ozel, K. Tutuncuoglu, S. Ulukus, and A. Yener, "Fundamental limits of energy harvesting communications," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 126–132, 2015.

[24] K. Tutuncuoglu, O. Ozel, A. Yener, and S. Ulukus, "The binary energy harvesting channel with a unit-sized battery," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4240–4256, 2017.

[25] B. Güler and A. Yener, "Energy-harvesting distributed machine learning," in *2021 IEEE Intl. Symp. Inf. Theory (ISIT)*. IEEE, 2021, pp. 320–325.

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

[27] L. Liu and W. Yu, "Massive connectivity with massive mimo—part i: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

[28] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," 2009.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.