Energy-Aware Analog Aggregation for Federated Learning with Redundant Data

Yuxuan Sun[*], Sheng Zhou[*], Deniz Gündüz[†]
[*]Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[†]Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, UK
Email: {sunyx15@mails., sheng.zhou@}tsinghua.edu.cn, d.gunduz@imperial.ac.uk

Abstract—**Federated learning (FL) enables workers to learn a model collaboratively by using their local data, with the help of a parameter server (PS) for global model aggregation. The high communication cost for periodic model updates and the non-independent and identically distributed (i.i.d.) data become major bottlenecks for FL. In this work, we consider analog aggregation to scale down the communication cost with respect to the number of workers, and introduce data redundancy to the system to deal with non-i.i.d. data. We propose an online energy-aware dynamic worker scheduling policy, which maximizes the average number of workers scheduled for gradient update at each iteration under a long-term energy constraint, and analyze its performance based on Lyapunov optimization. Experiments using MNIST dataset show that, for non-i.i.d. data, doubling data storage can improve the accuracy by 9.8% under a stringent energy budget, while the proposed policy can achieve close-to-optimal accuracy without violating the energy constraint.**

I. Introduction

With the rapid development of machine learning (ML) techniques, emerging applications, including virtual and augmented reality, Internet of things, autonomous driving and e-health services, are penetrating into human lives [1]. ML models for these applications are typically trained in central clouds. However, centralized training leads to high communication costs, and causes privacy concerns in applications that involve sensitive personal data.

Meanwhile, the end devices such as smart phones, vehicles and sensors and the infrastructures like base stations (BSs) and road side units are being equipped with more computing resources, enabling intensive computations at the network edge, namely *multi-access edge computing* [2]–[4]. With the help of edge intelligence and to address the privacy concerns, a distributed ML framework called *federated learning* (FL) has been proposed recently [5]–[7], where end devices, called *workers*, learn a shared ML model collaboratively using their local data, with the help of a central *parameter server (PS)* which aggregates the global model and coordinates the training process. Since the PS acquires the model update from each worker rather than their data, the privacy is preserved.

The high communication costs and non-independent and identically distributed (i.i.d.) data are the two major bottlenecks in FL [7]. According to [8], when using highly non-i.i.d. data for FL, the accuracy drops by 11% for MNIST[1] and 51% for CIFAR-10[2] as compared to that using i.i.d. data. They prove that the non-i.i.d. level, i.e., the difference between the local and global data distributions, is the root cause of the performance degradation. This problem is tackled by sharing publicly available i.i.d. data with the workers in [8], or workers sharing a limited portion of their data with the PS in [9].

The communication burden of FL mainly comes from the global model aggregation, which can be reduced by efficient scheduling and resource allocation [10]–[14], gradient quantization and sparsification [14]–[16], or via analog aggregation [15]–[17]. An analytical study on the convergence rate achieved by random, round robin and proportional fair scheduling policies is carried out in [10]. An energy-efficient bandwidth allocation and worker scheduling scheme is proposed in [11], minimizing the energy consumption while maximizing the fraction of workers scheduled. A more general resource constraint, including both communication and computing resources, is considered in [12], [13]. In [14], a hierarchical FL architecture is proposed, and the end-to-end latency is minimized by jointly considering model sparsification and the two-tier update interval. Quantization and error accumulation techniques are further considered in [15], [16] to reduce the communication cost.

Most papers on FL consider digital transmission for global model aggregation. However, the communication latency scales with the number of workers [17]. Observing that the PS is interested only in the *average* of local models rather than their individual values, a promising solution is to use *analog aggregation* which exploits the *waveform-superposition property* of a wireless multiple access channel (MAC) [15]–[17]. If the workers synchronize with each other and align the transmit power, the summation of local models can be carried out *over-the-air*. The tradeoff between signal-to-noise ratio (SNR) and the amount of exploited data is analyzed in [17], while gradient

[1]Modified National Institute of Standards and Technology database; a handwritten digit database. http://yann.lecun.com/exdb/mnist/

[2]Canadian Institute For Advanced Research; a collection of 60000 images in 10 categories. https://www.cs.toronto.edu/ kriz/cifar.html

(a) Architecture of FL, where redundancy comes from data exchange between workers.
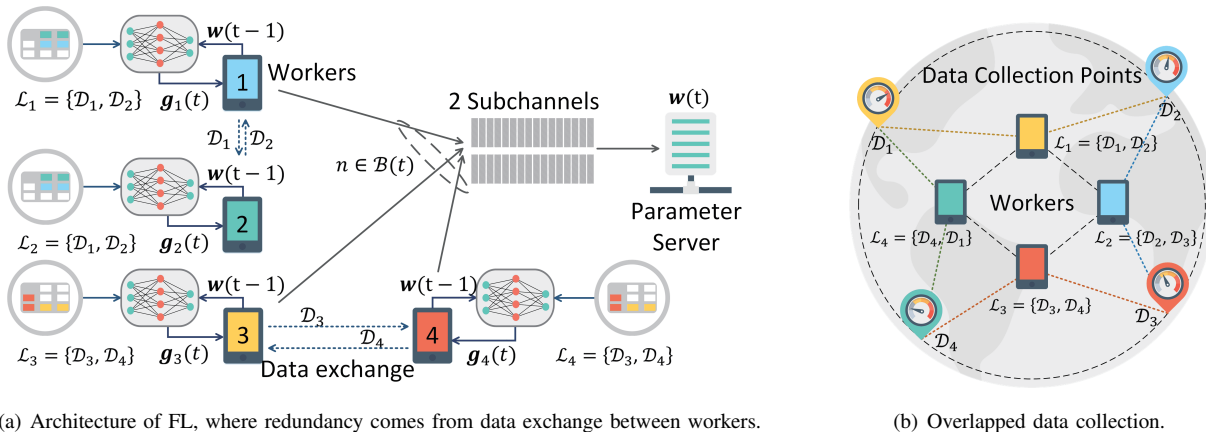
(b) Overlapped data collection.

Fig. 1. Illustration of an FL system and the acquisition of data redundancy.

compression and error accumulation are considered in [15], [16] to further improve the bandwidth efficiency of FL. While over-the-air aggregation requires channel state information (CSI) at the workers, it is shown in [18] that this requirement can be relaxed if the PS has multiple antennas.

Existing papers on analog aggregation mainly consider power allocation under specific channel models, and have not addressed the non-i.i.d. data. In this work, we consider *analog aggregation* for FL, where each worker has a *long-term energy budget*, and *data redundancy* is introduced to the system via data exchange or overlapped data collection. We propose an *energy-aware dynamic* worker scheduling policy, which maximizes the average weighted fraction of scheduled workers under average energy budget in an online fashion, and analyze its performance based on Lyapunov optimization. Experiments using MNIST dataset show that, data redundancy can bring significant accuracy improvement when data is non-i.i.d., while the proposed policy can smartly utilize the available energy to achieve close-to-optimal accuracy.

The rest of the paper is organized as follows. In Sec. II, we introduce the system model and problem formulation. The worker scheduling policy is proposed in Sec. III, along with its performance analysis. Experiment results are presented in Sec. IV, and the paper is finally summarized in Sec. V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Federated Learning Architecture

As shown in Fig. 1(a), we consider an FL system with a single PS and N homogeneous workers $\mathcal{N} = \{1, 2, \cdots, N\}$. To tackle non-i.i.d. data, we consider introducing *data redundancy* to the system via 1) *data exchange*, i.e., workers exchange their local data with neighboring workers they trust, or 2) *overlapped collection*, i.e., in IoT networks, the sensing area of sensors (workers) are overlapped with each other. For example, in Fig. 1(b), there are 4 data collection points, and each worker collects data from 2 points. As opposed to [8] and [9], this does not require sharing any data samples with the PS. We assume that there are K original datasets $\mathcal{D}_1, \cdots, \mathcal{D}_K$ generated by workers (in data exchange case) or data collection points (in

overlapped collection case), each has the same number of data samples, denoted by D. The global dataset is defined as $\mathcal{D} = \bigcup_{k=1,2,\cdots,K} \mathcal{D}_k$, with KD data samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{KD}\}$. For simplicity, we assume $K = N$ in the following.

The data redundancy of the system is denoted by r, indicating that each original dataset is stored at r different workers. The local dataset owned by worker n is denoted by $\mathcal{L}_n$, with $|\mathcal{L}_n| = rD$. One way to obtain redundancy r is to collect or exchange the original datasets in a cyclic manner: Let worker n store $\mathcal{L}_n = \bigcup_{n \in \mathcal{H}_n} \mathcal{D}_n$, where the set of indexes is $\mathcal{H}_n = \{H(n), \cdots, H(n + r - 1)\}$, with

$$H(n) = \begin{cases} n, & 1 \le n \le N, \\ n - N, & n > N. \end{cases} \tag{1}$$

Fig. 1(b) is an example of obtaining redundancy $r = 2$ in the data collection scenario, with $K = N = 4$.

The goal of FL is to minimize the global loss

$$\min_{\boldsymbol{w}} F(\boldsymbol{w}) \triangleq \frac{1}{ND} \sum_{i \in \mathcal{D}} f(\boldsymbol{w}, \boldsymbol{x}_i), \tag{2}$$

where $f(\boldsymbol{w}, \boldsymbol{x}_i)$ is a loss function designed for the FL task, and $\boldsymbol{w} \in \mathbb{R}^s$ is the parameter vector to be optimized.

In the t-th training round, the PS broadcasts the global parameter vector $\boldsymbol{w}(t-1)$ obtained in the last round to all the workers. We assume that the PS is a more capable node with sufficient energy resources (e.g., a BS); therefore the broadcast of the global parameter vector is error-free. Each worker randomly selects a fraction $\lambda_n(t)$ of data samples $\mathcal{L}_n(t) \subseteq \mathcal{L}_n$, with $|\mathcal{L}_n(t)| = \lambda_n(t)rD$, to evaluate its local gradient estimate $\boldsymbol{g}_n(t)$ as:

$$\boldsymbol{g}_n(t) = \frac{1}{\lambda_n(t)rD} \sum_{\boldsymbol{x}_i \in \mathcal{L}_n(t)} \nabla f(\boldsymbol{w}(t-1), \boldsymbol{x}_i). \tag{3}$$

Here we can let $\lambda_n(t) = \frac{1}{r}$, so that data redundancy *does not* bring additional computing workloads to the workers for training, but only increases the storage cost.

We define an indicator function $\beta_n(t)$, where $\beta_n(t) = 1$ if worker n is scheduled in the t-th round, and $\beta_n(t) = 0$

otherwise. We further define the set of workers scheduled in round t as $\mathcal{B}(t) = \{n \in \mathcal{N} : \beta_n(t) = 1\}$. The global parameter vector $\boldsymbol{w}(t)$ is updated according to

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) - \eta_t \frac{1}{|\mathcal{B}(t)|} \sum_{n \in \mathcal{B}(t)} \boldsymbol{g}_n(t), \qquad (4)$$

where η_t is the learning rate in round t.

B. Analog Aggregation

For the aggregation of the local gradients, we consider *analog transmission* via a wireless MAC with M sub-channels. If worker n is scheduled, its local gradient $\boldsymbol{g}_n(t)$ is evenly partitioned into M segments $\boldsymbol{g}_n(t) = [\boldsymbol{g}_{1,n}(t), \cdots, \boldsymbol{g}_{M,n}(t)]$, where, for $m = 1, \cdots, M$, $\boldsymbol{g}_{m,n}(t)$ is a vector with either $\lceil \frac{s}{M} \rceil$ or $\lfloor \frac{s}{M} \rfloor$ entries, and transmitted via sub-channel m.

In order to carry out the summation of the local gradients over-the-air, all the scheduled workers need to be synchronized and align their transmit power. Specifically, in round t, denote the power allocated to worker n within sub-channel m by $p_{m,n}(t)$, which satisfies

$$p_{m,n}(t) = \frac{\beta_n(t)\sigma_m(t)}{h_{m,n}(t)}, \qquad (5)$$

where $h_{m,n}(t)$ is the channel gain between worker n and the PS in sub-channel m, and $\sigma_m(t)$ is a power scalar to determine the received SNR. We assume that $h_{m,n}(t)$ remains constant within each round, but we do not limit $h_{m,n}(t)$ to any specific distribution. We also assume that each worker n has perfect knowledge of its current channel gains $h_{m,n}(t)$, $\forall m$. Within sub-channel m, each scheduled worker $n \in \mathcal{B}(t)$ transmits $p_{m,n}(t)\boldsymbol{g}_{m,n}(t)$ to the PS. The total communication latency in each round is $\lceil \frac{s}{M} \rceil$ times the symbol duration, regardless of the number of workers scheduled. We remark that, the consideration of sub-channels enables us to implement analog aggregation in the current digital transmit systems such as orthogonal frequency division multiplexing (OFDM) with minor changes [17]. However, we consider a worker-level schedule $\beta_n(t)$ rather than a sub-channel-worker-level schedule $\beta_{m,n}(t)$ in this work, so that the PS can receive the whole gradients of the scheduled workers, as shown in (4).

At the PS side, the received signal over sub-channel m can be written as

$$\boldsymbol{y}_m(t) = \sum_{n \in \mathcal{B}(t)} h_{m,n}(t)p_{m,n}(t)\boldsymbol{g}_{m,n}(t) + \boldsymbol{z}_m(t)$$
$$= \sigma_m(t) \sum_{n \in \mathcal{B}(t)} \boldsymbol{g}_{m,n}(t) + \boldsymbol{z}_m(t), \qquad (6)$$

where $\boldsymbol{z}_m(t)$ is an i.i.d. additive white Gaussian noise (AWGN) vector, with each entry following the standard normal distribution. The m-th segment of the global parameter vector, $\boldsymbol{w}_m(t)$, is updated according to

$$\boldsymbol{w}_m(t) = \boldsymbol{w}_m(t-1) - \frac{\eta_t}{|\mathcal{B}(t)|\,\sigma_m(t)}\boldsymbol{y}_m(t) = \boldsymbol{w}_m(t-1)$$
$$- \frac{\eta_t}{|\mathcal{B}(t)|} \sum_{n \in \mathcal{B}(t)} \boldsymbol{g}_{m,n}(t) - \frac{\eta_t}{|\mathcal{B}(t)|\,\sigma_m(t)}\boldsymbol{z}_m(t). \qquad (7)$$

And finally, the global parameters are concatenated into vector $\boldsymbol{w}(t) = [\boldsymbol{w}_1(t), \cdots, \boldsymbol{w}_M(t)]$.

C. Problem Formulation

Our objective is to minimize the global loss $F(\boldsymbol{w}(T-1))$ after T training rounds, by optimizing the worker schedule $\{\beta_n(t)\}$ and power allocation $\{\sigma_m(t), p_{m,n}(t)\}$. Meanwhile, we also want to explore how data redundancy r impacts the performance. The problem is formulated as:

$$\mathcal{P}1 : \min_{\{\beta_n(t),\ \sigma_m(t),\ p_{m,n}(t)\}} F(\boldsymbol{w}(T-1)) \qquad (8a)$$

$$\text{s.t.} \quad \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} \big\|p_{m,n}(t)\boldsymbol{g}_{m,n}(t)\big\|_2^2 \leq \bar{E}_n,\ \forall n, \qquad (8b)$$

$$p_{m,n}(t)h_{m,n}(t) = \sigma_m(t)\beta_n(t),\ \forall m, n, t, \qquad (8c)$$

$$\beta_n(t) \in \{0, 1\}, \forall n, t. \qquad (8d)$$

The first constraint (8b) states that the average energy consumed by each worker per training round cannot exceed budget $\bar{E}_n$, due to the battery limitation of wireless devices.[3] The second constraint (8c) states that all the scheduled workers align their power to enable over-the-air computation.

Since the loss function $F(\boldsymbol{w})$ is usually different for different kinds of machine learning tasks, and the evolution of the parameters during the training process is very complex, it is hard to express $F(\boldsymbol{w}(T-1))$ explicitly. Meanwhile, the convergence rate of distributed stochastic gradient descent is found to be positively correlated to the number of workers scheduled, as shown in [11] and references therein. Therefore, we consider an alternative optimization problem that maximizes the average weighted fraction of scheduled workers:

$$\max \frac{1}{T}\sum_{t=0}^{T-1} \frac{\gamma(t)\sum_{n=1}^{N}\beta_n(t)}{N}, \qquad (9)$$

where $\gamma(t)$ characterizes the importance of scheduling more workers in the t-th round.

We further fix $\sigma_m(t)$ to a predefined value, i.e., $\sigma_m(t) = \sigma$, $\forall t$. This assumption is based on the fact that, with analog aggregation, the convergence speed of the FL task is not very sensitive to the SNR or the average transmit power, according to [16]. Then $p_{m,n}(t) = \frac{\sigma\beta_n(t)}{h_{m,n}(t)}$ can be obtained from (8c) after deciding whether to schedule worker n or not. Finally, the energy consumption is given by

$$E_n(t) = \sum_{m=1}^{M} \left\|\frac{\sigma}{h_{m,n}(t)}\boldsymbol{g}_{m,n}(t)\right\|_2^2. \qquad (10)$$

The alternative problem can be formulated as

$$\mathcal{P}2 : \min_{\{\beta_n(t)\}} \frac{1}{T}\sum_{t=0}^{T-1} \gamma(t)\left(1 - \frac{\sum_{n=1}^{N}\beta_n(t)}{N}\right) \qquad (11a)$$

$$\text{s.t.} \quad \frac{1}{T}\sum_{t=0}^{T-1}\beta_n(t)E_n(t) \leq \bar{E}_n,\ \forall n, \qquad (11b)$$

$$\beta_n(t) \in \{0, 1\},\ \forall n, t. \qquad (11c)$$

---

[3]We unify each round to a unit time length, and use power and energy interchangeably in this paper without any ambiguity.

Algorithm 1 Energy-Aware Dynamic Scheduling Policy for FL via Analog Aggregation

---

1: **Initialization**: initialize global model $\boldsymbol{w}(-1)$, input N, σ, $\bar{E}_n$, $\gamma(t)$, V, and let $q_n(0) = q_{\min}$, $\forall n$.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3:      Broadcast $\boldsymbol{w}(t-1)$ to all the workers.   ▷ *By PS*
4:      Update $\boldsymbol{g}_n(t)$ from (3). ▷ *By Each worker, in parallel*
5:      Acquire channel gains $\{h_{m,n}\}$ for $m = 1, \cdots, M$, and calculate energy consumption $E_n(t)$ according to (10).
6:      Make scheduling decision:

$$\beta_n(t) = \begin{cases} 1, & \text{if } q_n(t)E_n(t) \leq \frac{V\gamma(t)}{N}, \\ 0, & \text{if } q_n(t)E_n(t) > \frac{V\gamma(t)}{N}. \end{cases} \quad (13)$$

7:      Update virtual queue $q_n(t)$ according to (15).
8:      Transmit $\frac{\sigma}{h_{m,n}(t)}\boldsymbol{g}_{m,n}(t)$ in sub-channel m, $\forall m$.
9:      Aggregate received signal $\boldsymbol{y}_m(t)$ according to (6), and update global model $\boldsymbol{w}(t)$ according to (7).   ▷ *By PS*
10: **end for**

III. ENERGY-AWARE WORKER SCHEDULING

The key challenge to solve $\mathcal{P}2$ is that, constraint (11b) is a long-term energy budget. However, in practice, the channel gains $h_{m,n}(t)$ and the power of the gradients $\|\boldsymbol{g}_{m,n}(t)\|_2^2$ cannot be acquired before the t-th round, and they may not be i.i.d. over time. Therefore, we design *online* worker scheduling policies in this section, and carry out performance analysis without assuming any specific distributions for the channel.

For any worker n, it is easy to see that its energy constraint (11b) and the scheduling decision $\beta_n(t)$ are independent of other workers. Then $\mathcal{P}2$ can be *equivalently* decoupled into N individual problems

$$\mathcal{P}3 : \min_{\{\beta_n(t)\}} \frac{1}{T}\sum_{t=0}^{T-1} \gamma(t)\frac{1-\beta_n(t)}{N} \quad (12a)$$

$$\text{s.t. } \frac{1}{T}\sum_{t=0}^{T-1} \beta_n(t)E_n(t) \leq \bar{E}_n, \quad (12b)$$

$$\beta_n(t) \in \{0,1\}, \ \forall t. \quad (12c)$$

The combination of the optimal solution of $\mathcal{P}3$ for all workers is the optimal solution of $\mathcal{P}2$. And by solving $\mathcal{P}3$, each worker can decide whether or not to send its gradient *individually*. In what follows, we design online solutions to $\mathcal{P}3$.

A. Myopic Scheduling for Short-term Fixed Energy Constraint

A simple way to handle the average energy constraint (12b) is to remove the long-term summation, and transform it to a short-term fixed energy constraint $\beta_n(t)E_n(t) \leq \bar{E}_n$, for $\forall t$. Then the *myopic scheduling* policy can be given by

$$\beta_n(t) = \begin{cases} 1, & E_n(t) \leq \bar{E}_n, \\ 0, & E_n(t) > \bar{E}_n. \end{cases} \quad (14)$$

In the t-th round, worker n acquires the current channel gains $\{h_{m,n}(t)\}$ and the powers of the gradient $\{\|\boldsymbol{g}_{m,n}(t)\|_2^2\}$ for $m = 1, 2, \cdots, M$, and calculates the required energy $E_n(t)$ to send it. If the required energy is no more than the budget $\bar{E}_n$, worker n is scheduled.

B. Energy-Aware Dynamic Scheduling for Long-term Average Energy Constraint

Although the myopic scheduling policy is simple and can satisfy the original long-term average energy budget, it actually introduces a much tighter energy constraint. In other words, the worker uses its energy in a more conservative manner, and is less likely to be scheduled compared with that allowed by the original energy budget (12b).

To schedule workers more efficiently, we propose an *energy-aware dynamic scheduling* policy, as shown in Algorithm 1. We construct a virtual queue $q_n(t)$ with $q_n(0) = q_{\min} \geq 0$, and its evolution is given by

$$q_n(t+1) = \max\{q_n(t) + \beta_n(t)E_n(t) - \bar{E}_n, q_{\min}\}. \quad (15)$$

The value of the virtual queue indicates the deficit between the current energy consumption and the budget.

As shown in Lines 3-4 in Algorithm 1, in each round, first, the PS broadcasts the up-to-date global parameter vector, and each worker runs a local gradient estimation step *in parallel* based on its local data. In Lines 5-6, each worker makes the scheduling decision $\beta_n(t)$ by comparing the *weighted energy* $q_n(t)E_n(t)$ and the *weighted utility* $\frac{V\gamma(t)}{N}$, where V is an adjustable weight parameter. The virtual queue $q_n(t)$ transforms the long-term energy budget into instantaneous energy constraint: if $q_n(t)$ is large, it is more likely that $q_n(t)E_n(t) > \frac{V\gamma(t)}{N}$, so that the worker tends to not send the gradient to save energy; and vice versa. Also, by introducing $q_n(t)$, $\beta_n(t)$ can be obtained without any future information, or other workers' states and decisions. Therefore, the proposed algorithm is *energy-aware*, *online* and *distributed*. As shown in Lines 7-9, all the workers then update their virtual queues, and the scheduled workers transmit their gradients to the PS synchronously. The global parameter vector is finally aggregated by the PS.

C. Performance Analysis

To analyze the performance of the proposed algorithm, we assume that $q_{\min} = 0$ in this subsection, and refer to the Lypunov optimization technique [19]. We consider a non-ergodic version of Lyapunov optimization, i.e., all the random variables can be non-i.i.d. over time; that is, 1) the distribution of the power of the gradient $\|\boldsymbol{g}_{m,n}(t)\|_2^2$ is unknown; 2) the channel and user mobility are not limited to specific models; 3) the total number of rounds T is finite.

Let $u_n(t) \triangleq \gamma(t)\frac{1-\beta_n(t)}{N}$. Define u_n^* as the optimal utility of $\mathcal{P}3$ achieved by the offline genie-aided solution, and $u_n^\dagger$ the average utility achieved by Algorithm 1. Let $u^* = \sum_{n=1}^{N} u_n^*$ be the optimal utility of $\mathcal{P}2$, and $u^\dagger = \sum_{n=1}^{N} u_n^\dagger$. By applying the energy-aware dynamic scheduling policy, we get the following theorem.

Theorem 1. *When $q_{min} = 0$, the average weighted fraction of scheduled workers achieved by Algorithm 1 satisfies:*

$$u^\dagger \leq u^* + \frac{T}{2V} \sum_{n=1}^{N} \alpha_n^2, \qquad (16)$$

and the total energy consumption of worker n is bounded by:

$$\sum_{t=0}^{T-1} \beta_n(t) E_n(t) \leq T\bar{E}_n + \sqrt{T^2 \alpha_n^2 + 2VT u_n^*}, \qquad (17)$$

where $\alpha_n = \max_t \left\{ \left| \beta_n(t) E_n(t) - \bar{E}_n \right| \right\}$.

Proof. See Appendix A. □

Theorem 1 shows that, the average utility and the total energy consumption achieved by the proposed energy-aware dynamic scheduling policy have deviation bounds, compared with the optimal genie-aided policy and the energy budget, respectively. Both deviations are positively correlated to the maximum energy deficit α_n, and can be traded-off by the weight parameter V.

D. Discussions

We remark that, according to (13), when $q_{\min} = 0$, the scheduling indicator $\beta_n(t) = 1$ if $q_n(t) = q_{\min}$. However, the energy cost $E_n(t)$ may be very high, leading to a large upper bound on the energy deficit α_n, and thus a large deviation from the optimal genie-aided solution. Moreover, the worker cannot be scheduled for many rounds afterwards, which in turn reduces the average utility. Therefore, although $q_{\min} = 0$ in the classical Lyapunov optimization, and the worst-case utility and energy consumption can be guaranteed, in the experiments, we set $q_{\min} > 0$, which is more reasonable.

In Algorithm 1, the weight parameter $\gamma(t)$ is also involved in balancing the utility and energy consumption. In practice, we can set $\gamma(t)$ as a relatively large value at the beginning of the training process, and decrease it over time, since 1) scheduling more workers at the initial training rounds helps the learning process converge faster; 2) the power of the gradients reduces over time, according to the experiments; 3) the battery level decreases as time goes by, so that we should use energy more and more conservatively.

IV. EXPERIMENTS

In this section, we evaluate the gain of data redundancy r and the performance of the proposed worker scheduling policy for a digit recognition FL task, using the MNIST dataset with 60000 training samples and 10000 test samples. We consider $N = 50$ workers, $M = 100$ sub-channels and $T = 100$ training rounds. We divide the dataset in either an i.i.d. or a non-i.i.d. fashion. For i.i.d. data partition, the data samples are randomly partitioned into N equal-size datasets $\mathcal{D}_1, \cdots, \mathcal{D}_N$. For the non-i.i.d case, the data samples are first sorted by the order of digits, and each dataset is formed by 1200 data samples from a single digit. Each worker stores r of these datasets in a cyclic manner according to (1), a total of $1200r$ samples. During each round, a fraction $\lambda_n(t) = \frac{1}{r}$ of data
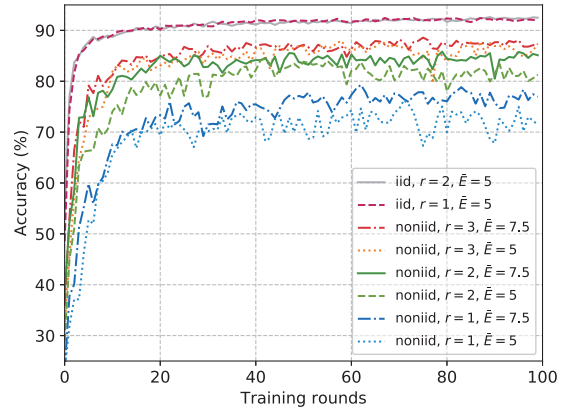

Fig. 2. The accuracy of the MLP under different data redundancies using the myopic scheduling policy.
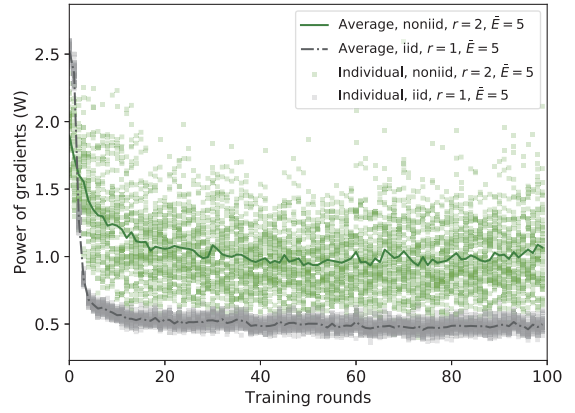

Fig. 3. Variation of the power of the gradients with training rounds.

samples are randomly chosen from the $1200r$ data samples worker n stores for training, so that the amount of computation for training is the same under different data redundancies.

We train a multilayer perceptron (MLP) with a 784-neuron input layer, a 64-neuron hidden layer, and a 10-neuron softmax output layer. The total number of parameters is 50890. Cross entropy is adopted as the loss function, and rectified linear unit (ReLU) activation is used. We set the learning rate η_t as 0.05, the dropout probability as 0.5, and use a momentum of 0.5. We consider Rayleigh fading channels from the workers to the PS, where the channel gain $h_{m,n}(t)$ follows the standard complex normal distribution, and each entry of the additive noise vector $\boldsymbol{z}_m(t)$ follows standard normal distribution. Also, the power scalar is set as $\sigma = 1$. For the energy-aware dynamic scheduling policy, we let $V = 1500$, and $q_{\min} = 0.3$.

In Fig. 2, we first explore the impact of data redundancy by evaluating the accuracy of the MLP using the myopic scheduling policy under both i.i.d. and non-i.i.d. data. The energy budget of each worker is set the same, with $\bar{E}_n = \bar{E} = 7.5$ J or 5 J. Overall, the accuracy with i.i.d. data outperforms that with non-i.i.d. data. When data is i.i.d., redundancy hardly brings any benefit to the system, as the workers can already use images of different digits to train the model even if $r = 1$. However, data redundancy can significantly improve the accuracy of the MLP under non-i.i.d. data. Specifically:
1) Given the energy budget, as the data redundancy increases,

the accuracy of the model increases with diminishing marginal gain. When $\bar{E} = 5$, an improvement of 9.8% is achieved by increasing redundancy from $r = 1$ to $r = 2$, and 6% from $r = 2$ to $r = 3$. 2) Increasing the energy budget helps to improve the performance of the MLP since more workers can be scheduled in each round, while the system is less sensitive to the energy constraint in the case of i.i.d. data. Comparing $\bar{E} = 5$ to $\bar{E} = 7.5$ with non-i.i.d. data, the accuracy improves by 5.5% when $r = 1$, by 3.8% when $r = 2$, but less than 1% when $r = 3$.

In Fig. 3, we plot the power of the gradients $||\boldsymbol{g}_n(t)||_2^2$ to guide the parameter design of the energy-aware dynamic scheduling policy. Each scatter point represents the power of the gradient of a worker in that round, and the lines are obtained by averaging these across workers, i.e., $\frac{1}{N}\sum_{n=1}^{N}||\boldsymbol{g}_n(t)||_2^2$. We find that, the power of the gradients reduces dramatically in the first 10-15 training rounds, and is quite stationary afterwards. Motivated by this observation, the weight parameter $\gamma(t)$ is set as:

$$\gamma(t) = \begin{cases} 2, & 0 \le t < 9, \\ 2 - 0.2(t-9), & 10 \le t < 15, \\ 1, & t \ge 15. \end{cases} \quad (18)$$
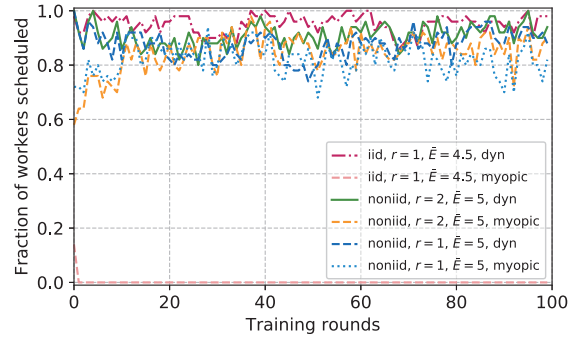
The performance of the proposed energy-aware dynamic scheduling policy is then presented in Fig. 4. In Fig. 4(a), we compare the accuracy of the dynamic scheduling policy with the myopic policy and an upperbound, which is achieved by setting $\bar{E} = \infty$, so that all the workers can be scheduled in each round. We can see that, under both i.i.d. and non-i.i.d. data distributions, the proposed algorithm can achieve close-to-optimal accuracy, while outperforming the myopic policy. Note that under i.i.d. data distribution, we further reduce the energy budget to $\bar{E} = 4.5$ J. Since the power of the gradient is high at the beginning of training, no worker can satisfy the energy constraint using the myopic policy. However, the dynamic scheduling policy enables workers to borrow energy from the future, so that workers can be scheduled. Fig. 4(b) and Fig. 4(c) present the fraction of workers $\frac{\sum_{n=1}^{N}\beta_n(t)}{N}$ scheduled in each round, and the maximum cumulative energy consumption $\max_{\{n \in \mathcal{N}\}} \sum_{\tau=0}^{t} \beta_n(\tau) E_n(\tau)$ over workers, respectively. By using the dynamic scheduling policy, more workers can be scheduled compared with the myopic policy, and the energy can be fully utilized. Specifically, with non-i.i.d. data, when $\bar{E} = 5$ J and data redundancy is $r = 2$, dynamic policy schedules an average of 90.9% workers in each round, while myopic policy can only schedule 84.6% workers. Moreover, the dynamic scheduling policy schedules more workers at the beginning while the energy is sufficient, so that the convergence of the MLP is accelerated.
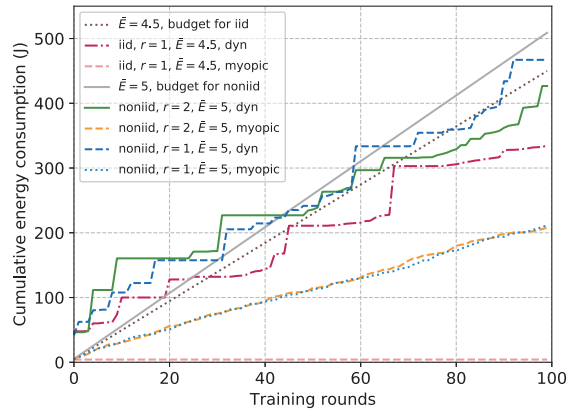
V. Conclusions

We have considered analog aggregation for FL over wireless channels and introduced data redundancy to the system to deal with non-i.i.d. data. We have proposed an energy-aware worker scheduling policy to maximize the weighted fraction
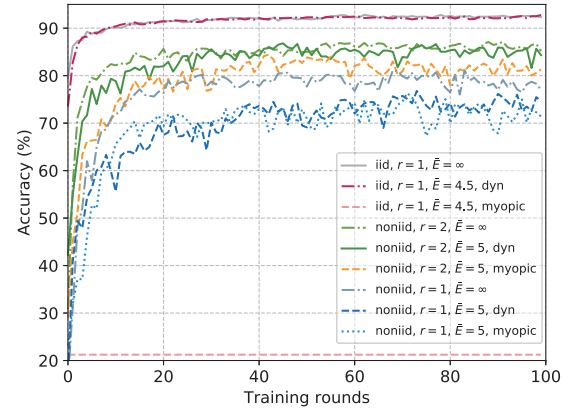
(a) Accuracy.

(b) Fraction of workers scheduled.

(c) Cumulative energy consumption.

Fig. 4. The performance of the proposed energy-aware dynamic scheduling policy, compared with the myopic scheduling policy and the upperbound.

of scheduled workers, which works in an online, distributed manner with performance guarantee. Experiments on the MNIST dataset have been carried out. Under our simulation settings, results show that for non-i.i.d. data, increasing data redundancy to $r = 2$ can improve the accuracy by 9.8% under a stringent energy budget. Further increases in redundancy lead to diminishing improvements in accuracy. We have also shown that, the proposed energy-aware dynamic scheduling policy can achieve close-to-optimal performance without violating the energy budget, and on average, schedule 6% more workers than a heuristic myopic policy. In the future, we plan to

quantify the impact of data redundancy, and consider gradient compression and error accumulation to further reduce the communication cost. Perfect synchronization among workers and accurate CSI are prerequisites in this work, while the impact of synchronization error and delayed CSI on training accuracy should also be studied.

APPENDIX A
PROOF OF THEOREM 1

Let $y_n(t) = \beta_n(t)E_n(t) - \bar{E}_n$. From (15), we have $y_n(t) \leq q_n(t+1) - q_n(t)$, $q_n^2(t+1) \leq (q_n(t) + y_n(t))^2$, $|q_n(t+1) - q_n(t)| \leq |y_n(t)|$, and $(q_n(t+1) - q_n(t))y_n(t) \leq y_n^2(t)$. When $q_{\min} = 0$, $q_n(0) = 0$, and thus

$$\sum_{t=0}^{T-1} y_n(t) = \sum_{t=0}^{T-1} \beta_n(t)E_n(t) - T\bar{E}_n \leq q_n(T). \tag{19}$$

Define the Lyapunov function as $L_n(t) \triangleq \frac{1}{2}q_n^2(t)$, and the one-slot Lyapunov drift as:

$$\Delta_n^{[1]}(t) \triangleq L_n(t+1) - L_n(t) = \frac{1}{2}q_n^2(t+1) - \frac{1}{2}q_n^2(t)$$
$$\leq \frac{1}{2}y_n^2(t) + q_n(t)y_n(t) \leq \frac{1}{2}\alpha_n^2 + q_n(t)y_n(t), \tag{20}$$

where $\alpha_n = \max_t \{|y_n(t)|\}$. The one-slot drift-plus-penalty function is given by:

$$\Delta_n^{[1]}(t) + Vu_n(t) \leq \frac{1}{2}\alpha_n^2 + q_n(t)y_n(t) + Vu_n(t). \tag{21}$$

The main idea of the proposed dynamic scheduling policy is to minimize the right-hand-side of (21). We have:

$$\min_{\beta_n(t)} q_n(t)y_n(t) + Vu_n(t) \tag{22}$$
$$\Leftrightarrow \min_{\beta_n(t)} \beta_n(t)\left(q_n(t)E_n(t) - \frac{V\gamma(t)}{N}\right) - q_n(t)\bar{E}_n + \frac{V\gamma(t)}{N}.$$

Since $\beta_n(t) \in \{0, 1\}$, the optimal solution of (22) is:

$$\beta_n(t) = \begin{cases} 1, & q_n(t)E_n(t) \leq \frac{V\gamma(t)}{N}, \\ 0, & q_n(t)E_n(t) > \frac{V\gamma(t)}{N}. \end{cases} \tag{23}$$

Define the T-slot drift as $\Delta_n^{[T]} \triangleq L_n(T) - L_n(1) = \frac{1}{2}q_n^2(T)$. Then the T-slot drift-plus-penalty function can be bounded by:

$$\Delta_n^{[T]} + V\sum_{t=0}^{T-1} u_n(t) \leq \sum_{t=0}^{T-1}\left(\frac{1}{2}\alpha_n^2 + q_n(t)y_n(t)\right)$$
$$+ V\sum_{t=0}^{T-1} u_n(t) \leq \frac{T}{2}\alpha_n^2 + \sum_{t=0}^{T-1}(q_n^*(t) - q_n(0))\, y_n^*(t) + VTu_n^*$$
$$\leq \frac{T}{2}\alpha_n^2 + \sum_{t=0}^{T-1} t\alpha_n^2 + VTu_n^* = \frac{T^2}{2}\alpha_n^2 + VTu_n^*, \tag{24}$$

where u_n^*, $q_n^*(t)$ and $y_n^*(t)$ are the optimal utility of $\mathcal{P}3$ obtained by the optimal genie-aided policy, and the corresponding value of queue and energy deficit. The inequality is obtained since the proposed algorithm minimizes $q_n(t)y_n(t) + Vu_n(t)$ in each round.

Since $\Delta_n^{[T]} \geq 0$ and $u_n(t) \geq 0$, from (19) and (24), we have

$$\sum_{t=0}^{T-1} \beta_n(t)E_n(t) - T\bar{E}_n \leq q_n(T) \leq \sqrt{T^2\alpha_n^2 + 2VTu_n^*},$$

$$u_n^\dagger \leq u_n^* + \frac{T}{2V}\alpha_n^2. \tag{25}$$

By summing (25) over $n = 1, \cdots, N$, we prove Theorem 1.

REFERENCES

[1] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," [Online] Available: https://arxiv.org/abs/1812.02858, Dec. 2018.

[2] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322-2358, 2017.

[3] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96-101, Jan.-Feb. 2018.

[4] S. Zhou, Y. Sun, Z. Jiang, and Z. Niu, "Exploiting moving intelligence: Delay-optimized computation offloading in vehicular fog networks," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 49-55, May 2019.

[5] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *NIPS Workshop on Private Multi-Party Machine Learning*, [Online] Available: https://arxiv.org/abs/1610.05492, Oct. 2016.

[6] K. Bonawitz, et al. "Towards federated learning at scale: System design," In *Conference on Systems and Machine Learning*, Stanford, CA, USA, Apr. 2019.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," [Online] Available: https://arxiv.org/abs/1908.07873, Aug. 2019.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," [Online] Available: https://arxiv.org/abs/1806.00582, Jun. 2018.

[9] N. Yoshida, T. Nishio, M. Morikura, K. Yamamoto, and R. Yonetani, "Hybrid-FL: Cooperative learning mechanism using non-iid data in wireless networks," [Online] Available: https://arxiv.org/abs/1905.07210, May 2019.

[10] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317-333, Jan. 2020.

[11] Q. Zeng, Y. Du, K. K. Leung, and K. Huang, "Energy-efficient radio resource allocation for federated edge learning," [Online] Available: https://arxiv.org/abs/1907.06040, Jul. 2019.

[12] S. Wang, et al. "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205-1221, Jun. 2019.

[13] N. H. Tran, W. Bao, A. Zomaya, Minh N.H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," *IEEE Conf. on Computer Commun. (INFOCOM)*, Paris, France, May 2019.

[14] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," [Online] Available: https://arxiv.org/abs/1909.02362, Sept. 2019.

[15] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, to appear, 2020.

[16] M. Mohammodi Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, to appear, 2020.

[17] G. Zhu, Y. Wang, and K. Huang, "Low-latency broadband analog aggregation for federated edge learning," *IEEE Trans. Wireless Commun.*, to appear, 2020.

[18] M. Mohammadi Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Ottawa, Canada, Nov. 2019.

[19] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.