This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

Federated Learning in Unreliable and Resource-Constrained Cellular Wireless Networks

Mohammad Salehi and Ekram Hossain, *FIEEE*

***Abstract*—With growth in the number of smart devices and advancements in their hardware, in recent years, data-driven machine learning techniques have drawn significant attention. However, due to privacy and communication issues, it is not possible to collect this data at a centralized location. Federated learning is a machine learning setting where the centralized location trains a learning model over remote devices. Federated learning algorithms cannot be employed in the real world scenarios unless they consider unreliable and resource-constrained nature of the wireless medium. In this paper, we propose a federated learning algorithm that is suitable for cellular wireless networks. We prove its convergence, and provide a sub-optimal scheduling policy that improves the convergence rate. We also study the effect of local computation steps and communication steps on the convergence of the proposed algorithm. We prove, in practice, federated learning algorithms may solve a different problem than the one that they have been employed for if the unreliability of wireless channels is neglected. Finally, through numerous experiments on real and synthetic datasets, we demonstrate the convergence of our proposed algorithm.**

***Index Terms*—Machine learning, federated learning, cellular wireless networks, success probability, signal-to-interference-plus-noise ratio (SINR), stochastic geometry, convergence analysis**

I. Introduction

A. Motivation

With the rapid growth in Internet-of-Things (IoT) applications and increase in the computational and storage power of smart devices, modern distributed networks generate a huge amount of data everyday [1]. Owing to this reason, data-driven machine learning techniques have gained significant attention in recent years. Currently, most of the existing machine learning techniques are centralized, i.e. they assume all data is available at a centralized location, where a central processor trains a powerful learning method on the data [2]. However, transferring data from user devices to the centralized location violates users privacy [3]. To cope with this issue, federated learning has been introduced where a learning model is trained over remote devices under the control of the centralized location, called server [1], [3], [4]. Specifically, in federated learning (FL), the server broadcasts the global model parameters to the remote devices. Each remote device uses its local dataset to update the global model, and then transmits the updated local model to the server. After aggregating the local models, the server updates the global model and repeats the whole procedure. As an example, consider the task of next-word prediction on mobile phones, where a language model predicts the most probable next word or phrase based on a small amount of user-generated preceding text. To maintain the users privacy, instead of transmitting the raw text data to the server and training a predictor at the server, we use federated learning [1], [5].

Although the above definition of federated learning seems similar to parallel optimization and distributed machine learning in datacenters, due to the following challenges, it needs to be treated separately: i) In federated learning, connections between the remote devices and the server[1] are unreliable and slow, ii) different devices in the network have different systems characteristics (systems heterogeneity), and iii) training data are not independently and identically distributed (statistical heterogeneity) [6], [7]. Since data is non-i.i.d. (not independently and identically distributed) across devices, all devices must participate in the learning process. However, due to the systems heterogeneity, which includes variable computation and communication capabilities at different devices, and limited amount of available resources such as bandwidth (e.g. number of resource blocks) for communication, full device participation at each round of communication is not possible. Moreover, in reality, transmission success probability is different for different devices, even when they all have the same hardware. Specifically, transmission success probability for devices that are located closer to the server is generally higher than that for devices that are located far. Thus, unless this issue is considered at the time of updating the global model, the updated global model will be biased towards cell center devices' local models.

B. Related Works

Training federated learning models in a wireless networking environment requires devices and the server exchange information via wireless transmissions. The existing works (e.g. in [3], [8], and references therein) aim to improve implementation of federated learning by optimizing resource allocation and/or reducing the communication requirements since communication is a key bottleneck [1]. To reduce the communication rounds of FL, three main approaches have been employed: quantization, sparsification, and local updates [9]. In this paper, we study local updates where, between any two aggregation steps in consecutive rounds, each device performs multiple local update steps. In this regard, [2], [10], [11] studied the convergence of communication-efficient

The authors are with the Department of Electrical and Computer Engineering at the University of Manitoba, Canada (emails: salehim@myumanitoba.ca, Ekram.Hossain@umanitoba.ca). **E. Hossain** is the corresponding author. The work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

[1]In a cellular wireless network, the base stations (BSs) can act as the servers.

FL with local updates for convex and non-convex problems. However, it was assumed that all of the devices participate in the aggregation step, which is obviously not possible when the number of available resource blocks is limited. To tackle this problem, in [6], [7], [12]–[15] at the beginning of each round, the server samples a subset of devices and allocates the available resource blocks to these devices. After performing local update steps, the BS aggregates the local models of the chosen (scheduled) devices and updates the global model. The works in [6], [7], [12]–[15] assumed that the BS successfully receives the local models of all the scheduled devices, and they designed the global model update step only based on the scheduling policy. However, in reality, not only is the success probability less than one, but also it is different for different devices. Clearly, to update the global model we need to include both scheduling policy and success probability.

To better understand, consider a scenario with two devices and two resource blocks, where both devices are scheduled for sharing their updated local models with the server. Assume that the success probability is 1 for *device one* and is 0.1 for *device two*. In this scenario, the server receives the local model of *device two* once in every ten rounds on average, while the local model of *device one* is always successfully received. Obviously, without considering this aspect of wireless communications, the global model is biased towards *device one* at the end of the learning process.

In the above context, we propose an FL algorithm that is suitable for unreliable and resource-constrained cellular wireless networks. Specifically, at the end of each FL round, our algorithm updates the global model by allocating different weights to different devices, where weight for each device depends on the scheduling policy and its transmission success probability.

To the best of our knowledge, only the work in [16] considered the effect of success probability on the convergence of FL. Similar to [17], [16] solved the FL problem using primal-dual optimization method. However, when strong duality is not guaranteed, this method may not be useful [13]. Also, in the analysis, the FL global objective is assumed to be a function of linear combination of model parameters and the input features. Thus, convergence analysis for this method cannot be extended to other machine learning techniques.

C. Contributions and Organization

The most common federated learning algorithm in the literature is FedAvg (federated averaging) [12], which lacks convergence analysis. Several works in the literature made steps towards analyzing convergence of FedAvg by modifying the algorithm. For example, [6] proposed a different averaging scheme for FedAvg, and derived the convergence rate of the modified algorithm. [13] proposed FedProx, which is a generalization of FedAvg obtained by adding a proximal term to the local objectives. Recently, [7] has proposed SCAFFOLD which improves FedAvg by adding a correction term, called client drift, to local updates. However, as we discussed in the previous subsection, these works and all other related works do not consider the effect of transmission success probability.

In the above context, the major contributions of this paper can be summarized as follows:

- We propose an FL algorithm that is suitable for unreliable and resource-constrained wireless systems. In particular, for an FL system in a cellular wireless network, the BS selects a subset of devices at each round and updates the global model based on their updated local models by allocating different weights to them. Weight for each device depends on the scheduling policy and its transmission success probability. We use stochastic geometry tools to approximately calculate the success probability for each device.
- Our proposed FL algorithm solves the FL problem in the primal domain. We prove that, for strongly convex and smooth problems, the algorithm converges on non-i.i.d data with rate $\mathcal{O}(\frac{1}{T})$.
- We study two difference scheduling (sampling) policies. We also provide a sub-optimal scheduling policy based on the derived convergence rate.
- We study the effect of number of computation steps and communication steps on the convergence rate.
- We show that the existing works, which do not include the transmission success probability in the global model update step (e.g. those in [6], [7], [12], [13]), will not be suitable for a wireless communication environment, since they may converge to the solution of a different FL problem when the success probabilities are different for different devices.
- We verify the convergence of our proposed FL algorithm by experimenting over real and synthetic datasets. We also compare our results with centralized full batch gradient descent which can be considered as a benchmark for our algorithm.

The organization of the rest of the paper is as follows. In Section II, we introduce the system model and propose our FL algorithm. Then, in Section III, we provide the convergence analysis of our FL algorithm. Further analysis of the proposed FL algorithm and comparison with related algorithms are provided in Section IV. In Section V, we present the simulation results. Finally, Section VI concludes the paper. A summary of the major notations used in the paper is given in Table I.

TABLE I
SUMMARY OF MAJOR NOTATIONS

Notation	Description
λ	Base station (BS) intensity (average number of BSs in a unit area)
N	Number of devices in each cell
M	Number of available resource blocks at each BS
F, F_k	Global loss function, local loss function at device k
w, w^k	Global model parameters, local model parameters at device k
p_k	Weight of k-th nearest device
q_k	Average number of allocated resource blocks to device k at a sampling step
U_k	Success probability of device k
K	Total number of FL rounds (iterations)
E	Number of local SGD steps during each round of FL
ℓ	Number of transmission attempts at each aggregation step

II. System Model, Assumptions, and Proposed FL Algorithm

A. Network Model

Consider a single-tier cellular network where the locations of the BSs follow a homogeneous Poisson point process (PPP) Φ of intensity λ[2]. Each BS serves the user devices that are located in its Voronoi cell, i.e. each device is associated to its nearest BS. In each cell, N devices are uniformly distributed; we use subscript k to denote the k-th nearest device to the BS. Thus, r_k denotes the distance between the serving BS and its k-th nearest device. Each BS allocates M resource blocks for the learning process, where $M \leq N$.

B. Federated Learning

In order to learn a statistical model from the distributed data across user devices, BS tries to solve the following distributed optimization problem:

$$\min_{w} \qquad F(w) = \sum_{k=1}^{N} p_k F_k(w), \tag{1}$$

where w is the learning model parameters. p_k is the weight of the k-th nearest device such that $p_k \geq 0$ and $\sum_{k=1}^{N} p_k = 1$. $F_k(w)$ also denotes the local loss function at device k; it is defined as

$$F_k(w) = \frac{1}{n_k} \sum_{x \in \mathcal{D}_k} \mathcal{L}(w, x), \tag{2}$$

where $\mathcal{D}_k$ is the local dataset at device k, and is non-i.i.d. across different devices. $n_k = |\mathcal{D}_k|$ denotes the number of samples in $\mathcal{D}_k$. Thus, in (1), we can set $p_k = \frac{n_k}{n}$, where $n = \sum_{k=1}^{N} n_k$. $\mathcal{L}(w, x)$ also represents the loss function for data sample x.

Since the information is distributed across multiple devices, a BS cannot directly solve (1). Therefore, the BS and the devices collaboratively learn the optimum model parameters $w^*(= \arg\min F(w))$ by following an iterative algorithm. Specifically, after initializing the model parameters at the BS at time 0, each iteration (round) of the algorithm comprises: 1) *Sampling and broadcast*, 2) *Local stochastic gradient descent (SGD)*, and 3) *Aggregation and averaging*. In the following, we discuss each of these steps in detail for an iteration that starts at time t. This iteration is also shown in Fig. 1.

Sampling and Broadcasting: Due to the limited number of available resource blocks, the BS first selects a group of devices at time t and then broadcasts its model parameters. In this paper, we consider two different sampling schemes.

In *Scheme I*, the BS uniformly selects M devices out of N devices without replacement, i.e. each device uses at most one resource block. Let us denote the set of the selected (scheduled) devices at time t by $\mathcal{S}_t$. For *Scheme I*, at sampling time t, the BS allocates q_k resource blocks to device k on average, where

$$q_k = \mathbb{E}\left[\sum_{m=1}^{M} \mathbf{1}\left(k \in \mathcal{S}_t(m)\right)\right] = \frac{M}{N}. \tag{3}$$

$\mathcal{S}_t(m)$ in the above equation denotes the index of the scheduled device at time t for resource block m, and $\mathbf{1}(.)$ is the indicator function.

In *Scheme II*, at time t for resource block m, the BS samples a device with replacement from N devices with probabilities $\{\hat{q}_k\}$. Thus, sampling is independently and identically distributed over m, and some devices may use more that one resource block. For *Scheme II*, we have

$$q_k = \mathbb{E}\left[\sum_{m=1}^{M} \mathbf{1}\left(k \in \mathcal{S}_t(m)\right)\right] = M\hat{q}_k. \tag{4}$$

Local SGD: After receiving the global model parameters from the BS at time t, the scheduled device k initializes its local model as $w_t^k = w_t$, where w_t is the global model parameters sent by the BS, and w_t^k is the local model parameters at device k. Then device k performs E steps of SGD and updates its local model at each step as follows:

$$v_{t+i+1}^k = w_{t+i}^k - \eta_{t+i} \nabla F_k(w_{t+i}^k; \xi_{t+i}^k), \qquad i = 0, ..., E-1, \tag{5}$$

where η_{t+i} is the learning rate, and ξ_{t+i}^k is a sample uniformly chosen from the local dataset at device k[3]. Moreover,

$$w_{t+i}^k = \begin{cases} w_t, & i = 0 \\ v_{t+i}^k, & i = 1, 2, ..., E-1 \end{cases}.$$

Aggregation and Averaging: At the end of E steps of stochastic gradient descent, the scheduled devices transmit $v_{t+E}^k - w_t$ to the BS using the allocated resource block(s). After collecting the local models, the BS updates the global model as

$$w_{t+E} = w_t + \sum_{k=1}^{N} \sum_{m=1}^{M} \frac{p_k}{q_k U_k} \mathbf{1}\left(k \in \mathcal{S}_t(m), \text{SINR}_{k,m} > \theta\right)\left(v_{t+E}^k - w_t\right), \tag{6}$$

where $\text{SINR}_{k,m}$ denotes the received signal-to-interference-plus-noise ratio (SINR) of device k over resource block m. θ denotes the SINR threshold, and U_k is defined as

$$U_k = \mathbb{E}\left[\mathbf{1}\left(\text{SINR}_{k,m} > \theta\right) \mid k \in \mathcal{S}_t(m)\right],$$

i.e. U_k denotes the success probability of device k given that it is scheduled to use resource block m.

According to (6), the BS updates the global model by allocating different weights to different devices. The weight for device k is a function of scheduling policy (q_k) and

[2]Spatial distributions of base stations and users affect the convergence rate of the algorithm through success probabilities. Real world deployment of BSs lies between perfectly regular and completely random. Thus, perfectly regular (such as hexagonal grid) and completely random (such as Poisson point process [PPP] model) spatial models provide bounds for success probabilities [18]. Due to the analytical tractability of PPP models, the existing literature (e.g. [19], [20], and references therein) mostly resort to PPP to model the cellular networks. Moreover, it has been observed that success probabilities for a variety of cellular network models and transmission techniques can be well approximated by a simple horizontal shift of the most tractable model, the PPP model [21]. Finally, it is worth mentioning that this model is in direct agreement with the 3rd generation partnership project (3GPP) models [22].

[3]In later sections, we will discuss about generalizing our results to batch sizes which are larger than one.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

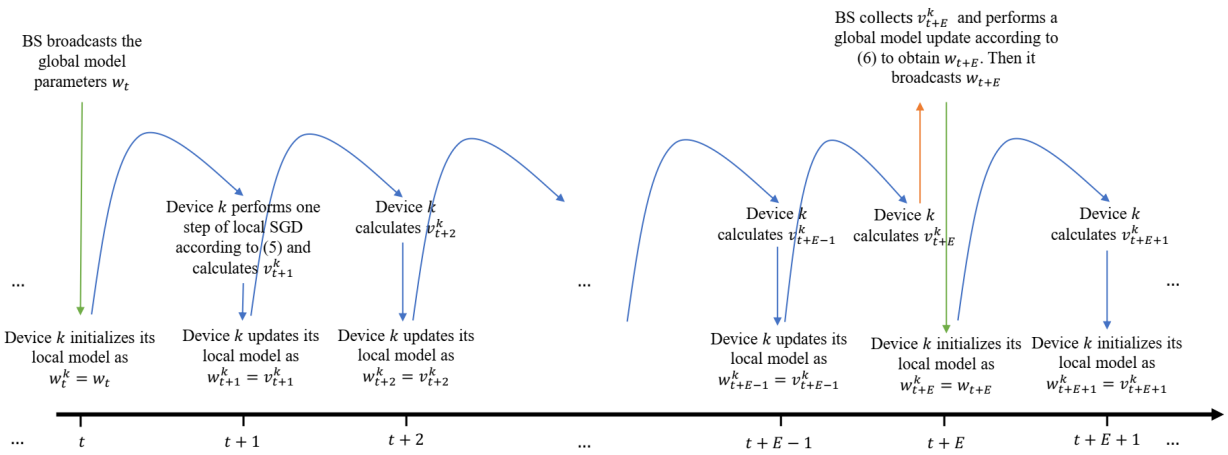


Fig. 1. One iteration (round) of the distributed learning algorithm starting at time t. Green arrows illustrate the broadcast step. Blue arrows correspond to the local SGD step, and red arrow represents the aggregation step.

transmission success probability (U_k) to compensate for systems heterogeneity and different communication capabilities. However, the averaging step in previous works, such as [6], [13], [15], is designed only based on the scheduling policy while assuming that success probabilities are all equal to one. As a result, the algorithms proposed in [6], [13], [15] (and other similar works) do not solve (1) in real world deployments, and this will be explained in Section IV.C.

Remark: From (5), we have

$$v_{t+E}^k - w_t = -\sum_{i=t}^{t+E-1} \eta_i \nabla F_k(w_i^k; \xi_i^k).$$

Thus, instead of transmitting $v_{t+E}^k - w_t$, device k can directly send the gradients (and also the learning rates). However, direct transmission of the gradients requires sending E times more parameters than $v_{t+E}^k - w_t$.

Based on the above discussion, we have summarized our proposed FL algorithm in **Algorithm 1**.

Algorithm 1: Proposed FL Algorithm

Input: K, E
Initialization: Global model parameters at the BS w_0
for $i = 0, 1, ..., K-1$ **do**
- **Step 1 (Sampling and Broadcast):** The BS samples a subset of devices and broadcasts the global model parameters w_{iE}.
- **Step 2 (Local SGD):** Each sampled device initializes its local model parameters with w_{iE}, and performs E steps of local SGD following (5) with $t = iE$.
- **Step 3 (Aggregation and Averaging):** Sampled devices transmit their updated local models to the BS. Then, the BS updates the global model parameters according to (6) and obtains $w_{(i+1)E}$.

return w_{KE}

Moreover, we can summarize the update rule of the proposed algorithm as

$$v_{t+1}^k = w_t^k - \eta_t \nabla F_k(w_t^k; \xi_t^k), \qquad \forall k \in \{1, ..., N\}, \quad (7)$$

where w_t^k is defined in (8). $\mathcal{I}_E = \{nE \mid n = 1, 2, ..., \}$ in (8) denotes the set of time indexes of the global synchronization steps after $t = 0$ [6].

C. Interference, SINR, and Success Probability

In this subsection, we study the success probability of device k at its associated BS, given that it is scheduled to transmit over resource block m. Due to the stationarity of the homogeneous PPP [23], we can consider the location of this BS as the origin of our coordination system.

Before studying the uplink success probability, it is worth mentioning that, in this paper, we assume downlink communication is always successful, i.e. all devices successfully receive the global model parameters. This is a valid assumption since the BS can transmit with more power compared to the devices; moreover, the BS can allocate more than one resource block for broadcasting the global model parameters. However, due to power constraint at user devices and limited number of allocated resource blocks to each scheduled device, we consider that uplink transmissions may not be successful at an aggregation step.

To increase the success probability during the aggregation steps, we assume that all scheduled devices transmit their local model parameters for ℓ times, and the BS employs selection combining, i.e. it uses the highest received SINR to recover the local model parameters transmitted over resource block m. Therefore,

$$U_k = \mathbb{E}\left[\mathbf{1}\left(\text{SINR}_{k,m} > \theta\right) \mid k \in \mathcal{S}_t(m)\right] =$$
$$\mathbb{E}[\mathbf{1}\left(\max\left\{\text{SINR}_{k,m}(1), ..., \text{SINR}_{k,m}(\ell)\right\} > \theta\right) \mid k \in \mathcal{S}_t(m)],$$

where $\text{SINR}_{k,m}(i)$ denotes the received SINR from device k over resource block m at i-th transmission attempt of an aggregation step.

Since the success event is identically distributed across different iterations (rounds) of the algorithm for device k,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

$$w_t^k = \begin{cases} w_0, & t = 0 \\ v_t^k, & t \notin \mathcal{I}_E \\ w_t = w_{t-E} + \sum_{k=1}^{N} \sum_{m=1}^{M} \frac{p_k}{q_k U_k} \mathbf{1}\left(k \in \mathcal{S}_{t-E}(m), \text{SINR}_{k,m} > \theta\right)\left(v_t^k - w_{t-E}\right), & t \in \mathcal{I}_E \end{cases} \tag{8}$$

$\text{SINR}_{k,m}(i)$ is not indexed by the iteration number. Moreover, at the aggregation step of each iteration, interference and fading are identically distributed across different resource blocks; thus, we can omit subscript m from $\text{SINR}_{k,m}(i)$ since its statistics do not depend on m. With this in mind, we can write

$$\text{SINR}_k(i) = \frac{h_k(i) r_k^{-\alpha}}{I(i) + \sigma^2} = \frac{h_k(i) r_k^{-\alpha}}{\sum_{x \in \Phi_\text{I}} h_x(i) \|x\|^{-\alpha} + \sigma^2}, \tag{9}$$

where σ^2 is the normalized noise power (noise power to the device transmit power). $I(i) = \sum_{x \in \Phi_\text{I}} h_x(i) \|x\|^{-\alpha}$ is the interference at i-th transmission attempt with Φ_I denoting the set of interferers. Here $h_x(i)$ denotes the small-scale fading gain between a device at location x and the BS at the origin. We consider Rayleigh fading, i.e. $h_x \sim \exp(1)$, and it is i.i.d. across i (different transmission attempts) and x (different locations). α denotes the path-loss exponent. Since the set of interferers remains the same during different transmission attempts of one aggregation step, the success events over one resource block during an aggregation step are temporally correlated.

It is also worth mentioning that our algorithm can also be employed with the results in [24] for maximal ratio combining. Moreover, we can easily extend our framework to more complex scenarios including uplink transmission with power control as in [25], multiple input multiple output antenna systems as in [26], [27], millimeter wave as in [28].

Lemma 1. *For the described network, with ℓ transmission attempts at each aggregation step, the success probability for scheduled device k is*

$$U_k \approx \sum_{i=1}^{\ell} \binom{\ell}{i} (-1)^{i+1} \exp\left\{ - i\theta\sigma^2 r_k^\alpha - 2\pi\lambda \right. \\ \left. \times \int_0^\infty \left(1 - \frac{1}{(1 + \theta r_k^\alpha x^{-\alpha})^i}\right)(1 - e^{-12/5\lambda\pi x^2}) x \mathrm{d}x \right\}.$$

Proof: See **Appendix A**. ∎

III. Convergence Analysis of the Proposed Federated Learning Algorithm

A. Convergence Rate

In this section, we prove the convergence of the proposed algorithm. In this regard, we first introduce additional notations and assumptions that are required to derive the convergence rate.

Following the same notation as in [6], for time t, we define

- $\bar{v}_t = \sum_{k=1}^{N} p_k v_t^k$,
- $\bar{w}_t = \sum_{k=1}^{N} p_k w_t^k$,
- $g_t = \sum_{k=1}^{N} p_k \nabla F_k(w_t^k; \xi_t^k)$,
- $\bar{g}_t = \sum_{k=1}^{N} p_k \nabla F_k(w_t^k)$, where $\nabla F_k(w_t^k) = \mathbb{E}_\xi\left[\nabla F_k(w_t^k; \xi)\right]$.

Thus, at $t \in \mathcal{I}_E$, $w_t^k = \bar{w}_t = w_t$.

We also make the following assumptions [6], [29], [30].

Assumption 1. *$F_1, ..., F_N$ are all μ-strongly convex, i.e. for $k \in \{1, ..., N\}$,*

$$F_k(w_2) \geq F_k(w_1) + \nabla F_k(w_1)^T (w_2 - w_1) + \frac{\mu}{2} \|w_2 - w_1\|^2, \quad \forall w_1, w_2.$$

Consequently, F is μ-strongly convex.

Assumption 2. *$F_1, ..., F_N$ are all L-smooth, i.e. for $k \in \{1, ..., N\}$,*

$$F_k(w_2) \leq F_k(w_1) + \nabla F_k(w_1)^T (w_2 - w_1) + \frac{L}{2} \|w_2 - w_1\|^2, \quad \forall w_1, w_2.$$

Consequently, F is L-smooth.

Assumption 3. *The variance of stochastic gradient at device k, $k \in \{1, ..., N\}$, is upper bounded by σ_k^2, i.e.*

$$\mathbb{E}_\xi\left[\|\nabla F_k(w; \xi) - \nabla F_k(w)\|^2\right] \leq \sigma_k^2, \qquad \forall w.$$

Assumption 4. *The second moment of the norm of the stochastic gradient is bounded at all devices. For all $k \in \{1, ..., N\}$,*

$$\mathbb{E}_\xi\left[\|\nabla F_k(w; \xi)\|^2\right] \leq G^2, \qquad \forall w.$$

We also define $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$, where F^* is the minimum global loss (objective function in (1)) and F_k^* is the minimum local loss at device k (objective function in (2)). $\Gamma \geq 0$, and increases as the heterogeneity (degree of non-i.i.d.) of the data distribution increases [6].

Proof: As mentioned previously, w^* minimizes the global loss, i.e. $F^* = F(w^*)$. Similarly, we define $w_k^* = \arg\min F_k(w)$. From the definition of Γ, we have

$$\begin{aligned} \Gamma &= F(w^*) - \sum_{k=1}^{N} p_k F_k(w_k^*) \\ &\overset{(a)}{=} \sum_{k=1}^{N} p_k F_k(w^*) - \sum_{k=1}^{N} p_k F_k(w_k^*) \\ &= \sum_{k=1}^{N} p_k \left(F_k(w^*) - F_k(w_k^*)\right) \overset{(b)}{\geq} 0, \end{aligned}$$

where (a) follows from the definition of the global loss function, and (b) is obtained since w_k^* minimizes F_k. When data is i.i.d. across devices, $F_1 = F_2 = \cdots = F_N = F$. Thus, $\Gamma = 0$ for i.i.d. data distribution. ∎

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

Theorem 1. *When* $\eta_t = \frac{2}{\mu(\gamma+t)}$ *with* $\gamma = \max\left\{8\frac{L}{\mu}, E\right\}$*, after the averaging step at time* T*, we have*

$$\mathbb{E}\left[F(w_T) - F^*\right] \leq \frac{L/\mu}{\gamma+T}\left[\frac{2}{\mu}\left(\sum_{k=1}^{N} p_k^2\sigma_k^2 + 6L\Gamma + 8(E-1)^2G^2 + 4E^2G^2B\right) + \frac{\mu\gamma}{2}\left\|w_0 - w^*\right\|^2\right],$$

where $B = \sum_{k=1}^{N} p_k\left(\frac{1}{q_k U_k} - 1\right)$ *for sampling Scheme I, and* $B = \sum_{k=1}^{N} p_k\left(\frac{1}{q_k U_k} - \frac{1}{M}\right)$ *for sampling Scheme II.* w_0 *denotes the initialized global model parameters.*

Proof: The result is obtained from

$$\mathbb{E}\left[\left\|\bar{w}_{t+1} - w^*\right\|^2\right] \leq (1-\mu\eta_t)\mathbb{E}\left[\left\|\bar{w}_t - w^*\right\|^2\right] + \eta_t^2\left(\underbrace{\sum_{k=1}^{N} p_k^2\sigma_k^2}_{\text{term I}} + \underbrace{6L\Gamma}_{\text{term II}} + \underbrace{8(E-1)^2G^2}_{\text{term III}} + \underbrace{4E^2G^2B}_{\text{term IV}}\right). \quad (10)$$

For details, see **Appendix B**. ∎

[6] established a convergence rate of $\mathcal{O}\left(\frac{1}{T}\right)$ based on **Assumptions 1 to 4**. According to **Theorem 1**, our proposed federated learning algorithm also converges with rate $\mathcal{O}\left(\frac{1}{T}\right)$ when **Assumptions 1 to 4** hold. Thus, unsuccessful transmissions do not affect the convergence significantly after proper adjustment of the averaging step.

In (10), **term I** is due to the fact that each device uses a mini-batch instead of the full batch to perform a local update, **term II** exists because data is non-i.i.d. across devices, **term III** stems from performing multiple ($E > 1$) local SGD steps between two aggregation steps which allows local models to move without the server's control in the direction of the local optimum instead of the global optimum, and **term IV** comes from the transmission unreliability and resource scarcity in cellular wireless networks. When **term I**, **term II**, **term III**, and **term IV** are zero, according to (10), the global model converges exponentially fast towards w^*. This is also the case in centralized full batch gradient descent. Thus, in Section V, we use centralized full batch gradient descent as a benchmark.

So far we have assumed that each device uses only one data sample at each local update step. In the following, we discuss using mini-batches with more than one data sample at the local update steps.

B. Mini-Batch Gradient Descent

When we use mini-batches with size b for local update steps, (7) changes to

$$v_{t+1}^k = w_t^k - \eta_t \nabla F_k(w_t^k; \{\xi_t^k\}), \qquad \forall k \in \{1, ..., N\}, \quad (11)$$

where

$$\nabla F_k(w_t^k; \{\xi_t^k\}) = \frac{1}{b}\sum_{\xi \in \mathcal{B}_t^k} \nabla F_k(w_t^k; \xi)$$

is the estimated gradient at device k at time t using samples in the mini-batch $\mathcal{B}_t^k = \{\xi_t^k\}$. For device k and $\forall w$, we have

$$\begin{aligned}
&\mathbb{E}_{\{\xi\}}\left[\left\|\nabla F_k(w; \{\xi\}) - \nabla F_k(w)\right\|^2\right] \\
&= \mathbb{E}_{\{\xi\}}\left[\left\|\sum_{\xi \in \mathcal{B}^k}\frac{1}{b}\left(\nabla F_k(w;\xi) - \nabla F_k(w)\right)\right\|^2\right] \\
&\leq \mathbb{E}_{\{\xi\}}\left[\sum_{\xi \in \mathcal{B}^k}\frac{1}{b}\left\|\left(\nabla F_k(w;\xi) - \nabla F_k(w)\right)\right\|^2\right] \\
&\overset{(a)}{\leq} \sigma_k^2,
\end{aligned}$$

where (a) is obtained from **Assumption 3**. Similarly, from **Assumption 4**, we have

$$\mathbb{E}_{\{\xi\}}\left[\left\|\nabla F_k(w; \{\xi\})\right\|^2\right] \leq G^2, \qquad \forall w.$$

Therefore, **Theorem 1** holds for any batch size.

IV. Further Analysis and Comparison

In this section, we provide a further discussion on the proposed federated learning algorithm using **Theorem 1**. Specifically, we first study the impact of number of iterations (rounds), number of local update steps at each round, and number of transmission attempts at each aggregation step in each round. Then we find a sub-optimal scheduling policy that improves the convergence rate. Finally, we compare our algorithm with the works that do not consider the transmission success probability and prove its significance.

A. Effects of Number of Computation and Communication Steps

To study the impact of number of rounds (K[4]), number of local update steps during each round (E), and number of transmission attempts at each aggregation step in each round (ℓ), we use a simpler form of **Theorem 1**.

Corollary 1. *When* $\eta_t = \frac{2}{\mu(\gamma+t)}$ *with* $\gamma = \max\left\{8\frac{L}{\mu}, E\right\}$*, after the averaging step at the* K*-th round, we have*

$$\mathbb{E}\left[F(w_T) - F^*\right] \leq \frac{L/\mu}{KE}\left[\frac{2}{\mu}\left(\sum_{k=1}^{N} p_k^2\sigma_k^2 + 6L\Gamma + 8(E-1)^2G^2 + 4E^2G^2B\right) + \frac{\mu\gamma}{2}\left\|w_0 - w^*\right\|^2\right],$$

where $T = KE$ *and* B *is defined in* ***Theorem 1****.*

When $T = KE$ is fixed, the right hand side of **Corollary 1** is minimum when $K = T$ and $E = 1$ because we only need to minimize $8(E-1)^2G^2 + 4E^2G^2B$. Therefore, the gap between the global model and the global optimum model is minimized at time T if the server updates the global model after every local update step.

[4]According to the definition, $K = T/E$ in **Theorem 1**.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

By taking the derivative of the right hand side with respect to E when K (number of FL rounds) is fixed, we observe that the upper bound in **Corollary 1** first decreases and then increases; thus, an optimal value for E exists. In fact this can be easily understood by considering that increasing E at first allows each device to move further in the direction of the optimum model parameters; thus, improves the convergence. However, when E is set to a large value, each device moves towards its local optimum model instead of the global optimum.

When E is fixed, it is obvious that the upper bound in **Corollary 1** is a decreasing function of K and ℓ, i.e. the the gap between the global model and the global optimum model decreases as we increase the number of communications. However, increasing K decreases the gap more than increasing ℓ. To prove this statement, we denote the success probability after ℓ transmission attempts by $U_k^{(\ell)}$. According to (A.1),

$$\begin{aligned} U_k^{(\ell)} &= 1-\mathbb{E}_{\Phi_{\mathrm{I}}}\left[\left(1-e^{-\theta\sigma^2 r_k^{\alpha}}\prod_{x\in\Phi_{\mathrm{I}}}\frac{1}{1+\theta r_k^{\alpha}\|x\|^{-\alpha}}\right)^{\ell}\right] \\ &\overset{(a)}{\leq} 1-\mathbb{E}_{\Phi_{\mathrm{I}}}\left[1-\ell e^{-\theta\sigma^2 r_k^{\alpha}}\prod_{x\in\Phi_{\mathrm{I}}}\frac{1}{1+\theta r_k^{\alpha}\|x\|^{-\alpha}}\right] \\ &\leq \ell U_k^{(1)}, \end{aligned} \tag{12}$$

where, in (a), we have used Bernoulli's inequality [31]. When we increase K to ℓK, $\ell > 1$, from the upper bound in **Corollary 1** we have

$$\begin{aligned} &\frac{L/\mu}{\ell KE}\left[\textbf{Constant}_1+\textbf{Constant}_2\times\sum_{k=1}^{N}\frac{p_k}{q_k U_k^{(1)}}\right] \\ &\leq \frac{L/\mu}{KE}\left[\textbf{Constant}_1+\textbf{Constant}_2\times\sum_{k=1}^{N}\frac{p_k}{\ell q_k U_k^{(1)}}\right] \\ &\overset{(a)}{\leq} \frac{L/\mu}{KE}\left[\textbf{Constant}_1+\textbf{Constant}_2\times\sum_{k=1}^{N}\frac{p_k}{q_k U_k^{(\ell)}}\right], \end{aligned}$$

where we have used (12) in (a). Although increasing K is more powerful compared to increasing ℓ in decreasing the gap, it consumes more resources since, during each round of the learning process, the scheduled devices perform E steps of local update.

B. Scheduling Policy

The optimal scheduling policy minimizes the global loss function, and can be obtained from the solution to the following optimization problem:

$$\begin{aligned} \underset{\{q_k\}}{\text{minimize}} \quad & \mathbb{E}\left[F(w_T)\right] \\ \text{subject to} \quad & \sum_{k=1}^{N} q_k = M, \\ & q_k > 0, \qquad \forall k\in\{1,...,N\}, \end{aligned} \tag{13}$$

where the first condition is due to the fact that the BS allocates M resource blocks for the learning process.

To solve (13), we need to understand how scheduling affects the global loss for a given w (global model parameters). Since in general this is impossible [2], we use the convergence bound in **Theorem 1** to approximately solve (13). Thus, we can obtain a sub-optimal solution to (13) by solving

$$\begin{aligned} \underset{\{q_k\}}{\text{minimize}} \quad & \sum_{k=1}^{N}\frac{p_k}{U_k}q_k^{-1} \\ \text{subject to} \quad & \sum_{k=1}^{N} q_k = M, \\ & q_k > 0, \qquad \forall k\in\{1,...,N\}. \end{aligned} \tag{14}$$

Now, consider the following optimization

$$\begin{aligned} \underset{\{q_k\}}{\text{minimize}} \quad & \sum_{k=1}^{N}\frac{p_k}{U_k}q_k^{-1} \\ \text{subject to} \quad & \sum_{k=1}^{N} q_k \leq M, \\ & q_k > 0, \qquad \forall k\in\{1,...,N\}. \end{aligned} \tag{15}$$

By contradiction, one can easily prove that solution to (15), denoted by $q_1^*,...,q_N^*$, satisfies $\sum_{k=1}^{N} q_k^* = M$. Thus, (15) is equivalent to (14), i.e. $q_1^*,...,q_N^*$ is also the solution to (14). (15) is a geometric program, and it can be transformed to a convex problem [32]. Therefore, we can find $q_1^*,...,q_N^*$ using CVX or any other convex solver [33]. Finally, the (sub-) optimal[5] policy for *Scheme II* is achieved when $\hat{q}_k = \frac{q_k^*}{M}$. Note that the optimal policy for *Scheme II* does not necessarily perform better than *Scheme I* since convergence rate and the upper bound in **Theorem 1** for *Scheme I* are different from *Scheme II*. In fact, it is only guaranteed that, with optimal policy for *Scheme II*, the upper bound in **Theorem 1** performs better compared to other *Scheme II* scheduling policies.

Finally, it is worth mentioning that, according to **Lemma 3**, the optimal scheduling policy minimizes the variance of the updated global model at each averaging step.

C. Comparison with the Existing Works

Existing works, such as [6], [7], [12], [13], assume that the BS always successfully recovers the transmitted information of the scheduled devices. For example, [6], [13], [15] use sampling *Scheme II* with probabilities $\{p_k\}$ besides the following averaging approach:

$$w_t = \frac{1}{M}\sum_{k=1}^{N}\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}_{t-E}(m)\right)v_t^k \tag{16}$$

at $t\in\mathcal{I}_E$. To study this averaging approach when unsuccessful transmission probability is greater than zero and varies across

[5]For the rest of the paper, by optimal scheduling policy we mean the scheduling policy that minimizes the convergence bound in **Theorem 1**. This scheduling policy is optimal for (14) and sub-optimal for (13).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

the devices, we modify it as

$$w_t = \begin{cases} w_{t-E}, \\ \text{if } \sum_{k=1}^{N}\sum_{m=1}^{M} \mathbf{1}\left(k \in \mathcal{S}_{t-E}(m), \text{SINR}_{k,m} > \theta\right) = 0 \\ \frac{1}{M}\sum_{k=1}^{N}\sum_{m=1}^{M} \mathbf{1}\left(k \in \mathcal{S}_{t-E}(m), \text{SINR}_{k,m} > \theta\right) v_t^k, \\ \text{otherwise} \end{cases} . \tag{17}$$

In the following, we prove that this averaging scheme does not solve (1) when unsuccessful transmission probability is greater than zero and varies across the devices. In this regard, we focus on the following averaging approach:

$$w_t = w_{t-E} + \sum_{k=1}^{N}\sum_{m=1}^{M} \frac{\mathbf{1}\left(k \in \mathcal{S}_{t-E}(m), \text{SINR}_{k,m} > \theta\right)}{\sum_{k=1}^{N}\sum_{m=1}^{M} \mathbf{1}\left(k \in \mathcal{S}_{t-E}(m), \text{SINR}_{k,m} > \theta\right)} \left(v_t^k - w_{t-E}\right), \quad t \in \mathcal{I}_E, \tag{18}$$

where we define $\frac{0}{0} = 0$. It is straightforward to show that, when $M = 1$, (18) is same as (17).

In **Appendix C**, we prove that by using (18) for updating the global model parameters at $t \in \mathcal{I}_E$, instead of (1), the algorithm converges to the solution to the following problem:

$$\min_{w} \qquad \hat{F}(w) = \sum_{k=1}^{N} p_k U_k F_k(w). \tag{19}$$

Since (18) and (17) are equal for $M = 1$, we can conclude that the employed learning approach by [6], [13], [15] do not solve (1) when U_k is less than one and is different for different devices. In fact, one can easily understand that, in this case, (17) is biased towards devices with higher success probabilities.

Finally, it is worth mentioning that (16) can be considered as a special case of (6), i.e. (6) simplifies to (16) when all transmissions are successful ($\text{SINR}_{k,m} > \theta$ and $U_k = 1$, $\forall k, m$).

V. Simulation Results and Discussion

Cellular Network: BS intensity λ is 0.001 (points/area). We assume in each cell there are $N = 100$ user devices and each BS allocates only $M = 20$ resource blocks for the distributed learning process. We also set $\sigma^2 = 10^{-4}$, $\alpha = 4$, and $\theta = -15\,\text{dB}$.

Datasets: We evaluate our proposed federated learning on both real and synthetic datasets. For real data, we use MNIST [34] and distribute our data similar to [6] in a non-i.i.d. fashion. Specifically, each sample in MNIST dataset is a 28×28 image of a handwritten digit between 0 to 9. We distribute the dataset samples such that each device has samples of only two digits and the number of samples at different devices is different. For synthetic data, we follow the same setup as in [6], [13], [35]. In this regard, we generate samples at device k, denoted by (X_k, Y_k), using $y = \text{argmax}\left(\text{softmax}\left(W_k x + b_k\right)\right)$, where $W_k \in \mathbb{R}^{10\times 60}$ and $b_k \in \mathbb{R}^{10}$. Each element in W_k and b_k is a realization of $\mathcal{N}(\mu_k, 1)$, where $\mu_k \sim \mathcal{N}(0, \tilde{\alpha})$. Thus, $\tilde{\alpha}$ controls the degree of difference between local models at different devices[6] [13]. Moreover, $x \in \mathbb{R}^{60}$. The j-th element (feature) in x is drawn from $\mathcal{N}(v_{k,j}, j^{-1.2})$, where $v_{k,j} \sim \mathcal{N}(B_k, 1)$[7] and $B_k \sim \mathcal{N}(0, \tilde{\beta})$. Thus, $\tilde{\beta}$ controls the data heterogeneity. Finally, the number of data samples at device k, denoted by n_k, follows a power law distribution. In this paper, we set $\tilde{\alpha} = 1$ and $\tilde{\beta} = 1$.

Model: To examine the performance of the proposed FL algorithm on the discussed datasets, we use a three layer neural network with 300 hidden units at each hidden layer. For MNIST datatset, input of the model is a flattened 784-dimensional image. For the synthetic dataset, an input of the model has 60 dimensions. We use ReLU (rectified linear unit) activation function for the hidden layers and softmax in the output layer. At the beginning of round k, $k \in \{0, 1, 2, ..., K-1\}$, we set the learning rate as $\eta_t = \frac{\eta_0}{1+k}$, where $kE \leq t < (k+1)E$. We measure the performance of our model with regularized cross-entropy loss, where we use ℓ_2-norm regularization with regularization parameter 10^{-4}. Therefore, the local loss at device k is computed by

$$F_k(w) = \frac{1}{n_k} \sum_{(x,y)\in\mathcal{D}_k} \text{CrossEntropy}(f(w;x), y) + 10^{-4}\|w\|_2^2,$$

where $f(w;x)$ denotes the neural network prediction for input x [6]. The global loss is a weighted average of the local losses as given in (1).

Benchmark: When $M = N$ and $U_k = 1$, $\forall k\{1, ..., N\}$, the BS receives the local model parameters of all devices successfully. Under this assumption, if all devices use their full batch and $E = 1$, the distributed learning algorithm is equivalent to the centralized full batch gradient descent. This scenario can be regarded as a benchmark for our proposed algorithm. In Fig. 2, performance of the centralized full gradient descent with $\eta_0 = 1$ is shown in terms of global loss and accuracy over both MNIST and synthetic datasets. Accuracy is defined as the percentage of correct predictions.

Training: To train the local models at local SGD steps, for MNIST dataset, we use batch sizes of 64 and set $\eta_0 = 1$. For the synthetic dataset, we use batch sizes of 25 and set $\eta_0 = 0.1$. Since we use a smaller learning rate for the synthetic dataset, we consider higher values for E at the time of simulation. Our codes are available at [36].

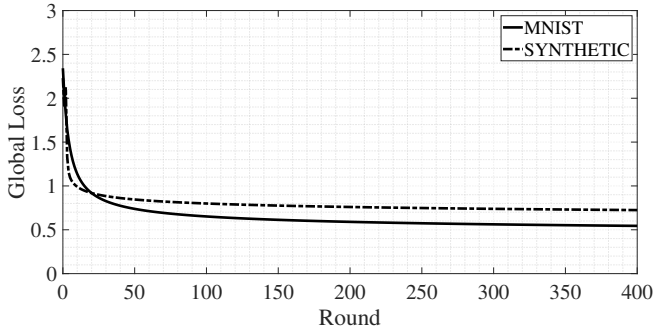
Results[8]: In Figs. 3(a) and (b), we measure the performance of the proposed FL algorithm in terms of global loss and accuracy on MNIST dataset; Fig. 3(c) and (d) also illustrate the performance on the synthetic dataset. For *Scheme II*, we show the results with $\hat{q}_k = \frac{1}{N}$, i.e. uniform selection and $\hat{q}_k = \frac{q_k^*}{M}$, i.e. (sub-) optimal selection. For MNIST dataset, the objective function of (15) at $q_1 = q_2 = ... = q_N = \frac{M}{N}$ is 7.09 and at the optimal point $q_1^*, q_2^*, ..., q_N^*$ is 6.48. For the synthetic dataset, these values are 6.89 and 3.66, respectively. Thus, optimal scheduling has more effect on the synthetic dataset which can

[6]When $\tilde{\alpha} = 0$, elements of W_k and b_k at all devices are drawn from the same distribution, i.e. $\mathcal{N}(0, 1)$.

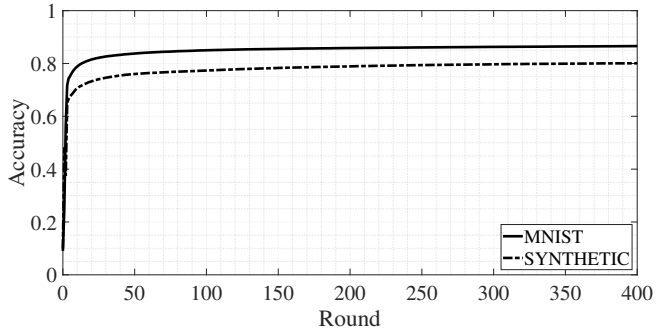
[7]$v_{k,j}$ remains the same across j-th feature of different samples of device k.

[8]To make a fair comparison, we have illustrated the average performance over multiple trials.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

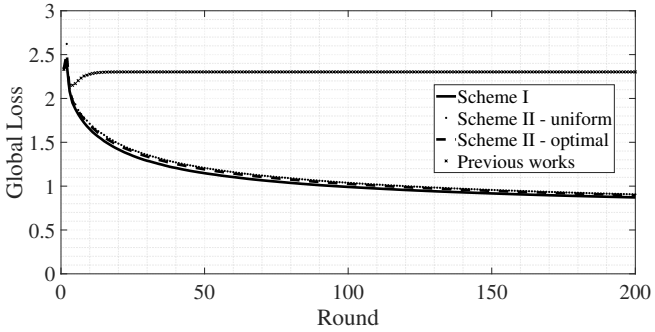


(a) Global loss.

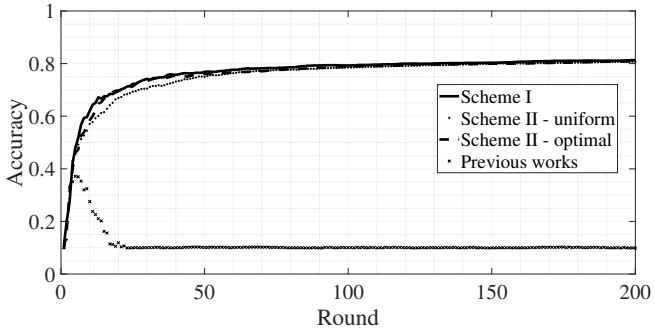


(b) Accuracy.

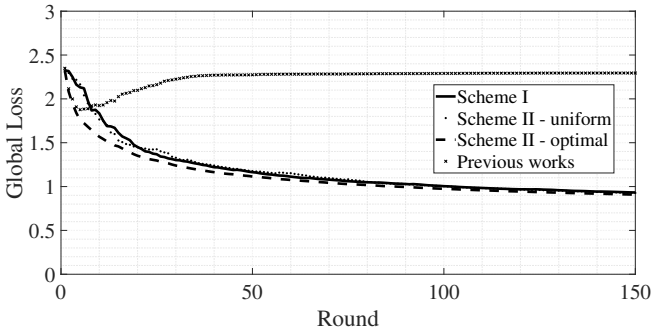
Fig. 2. Performance of the benchmark algorithm over MNIST and Synthetic datasets.



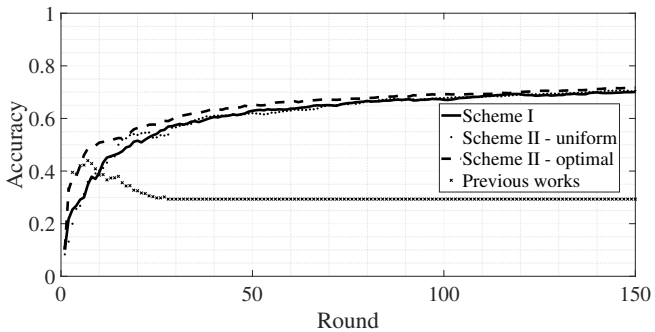
(a) MNIST.



(b) MNIST.



(c) Sythetic.



(d) Synthetic.

Fig. 3. In (a) and (b), $\ell = 2$ and $E = 1$. In (c) and (d), $\ell = 2$ and $E = 20$.

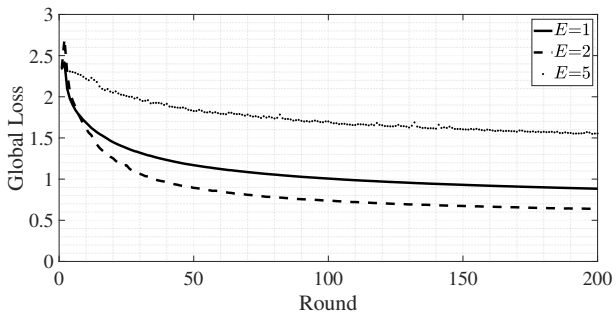
also be understood from Fig. 3. By calculating B in **Theorem 1** for *Scheme I* and *Scheme II* with optimal scheduling, we expect that *Scheme I* performs better on MNIST dataset and *Scheme II* performs better on the synthetic dataset. This is also illustrated is Fig. 3 and is in compliance with our discussion in **Section IV.B**. Moreover, as explained in **Section IV.C**, previous works, such as [6], [13], [15], cannot be used for wireless networks since they do not include unsuccessful transmission probability. This is also shown in Fig. 3, where we illustrate the performance of the introduced setup in **Section IV.C**.

We provide more simulation results in Figs. 4, 5, 6, and 7 where we study the effects of K, ℓ, and E on the performance of our algorithm. For *Scheme I* sampling, we show the results in Figs. 4 and 6. For *Scheme II* sampling with $\hat{q}_k = \frac{1}{N}$, we show the results in Figs. 5 and 7. Note that, when $\hat{q}_k = \frac{1}{N}$ in *Scheme II*, both schemes select users uniformly; however, *Scheme I* selects without replacement, while replacement is allowed in *Scheme II*. We see that *Scheme I* performs better than *Scheme II* with uniform selection. This can be also understood from **Theorem 1**. Finally, it is worth mentioning that, with appropriate tuning of the hyper-parameters (K, ℓ, and E), we can reach the benchmarked performance levels.

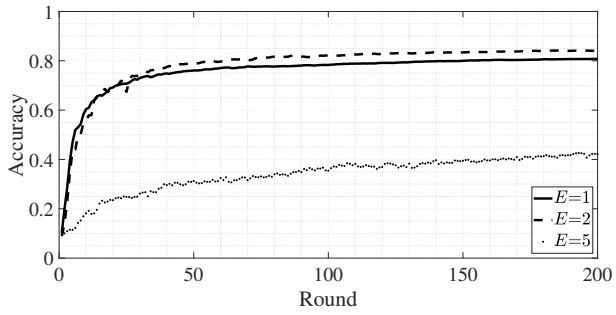
VI. Conclusion

Federated learning algorithms cannot be employed in the real world scenarios unless they consider the scarcity of radio resources and unreliability of wireless transmissions. In this regard, in this paper, we have proposed a federated learning algorithm that is tailored for unreliable and resource-constrained

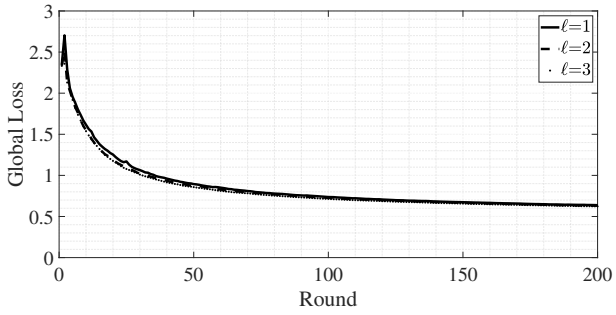
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications



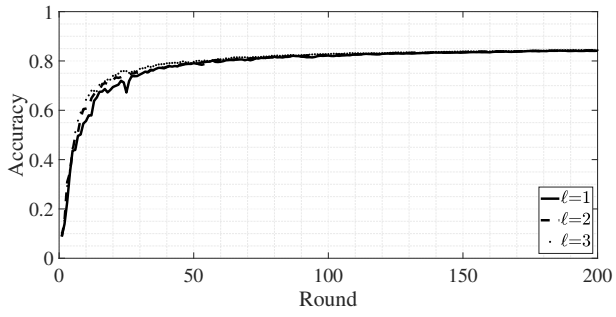
(a) Global loss.



(b) Accuracy.

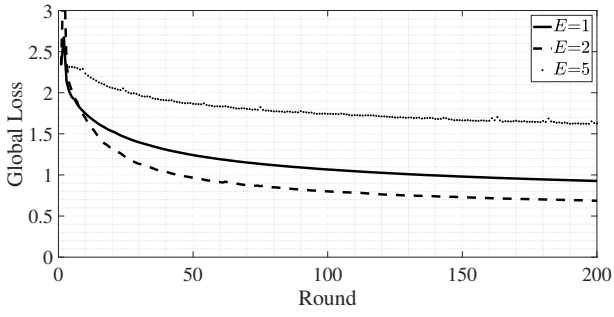


(c) Global loss.

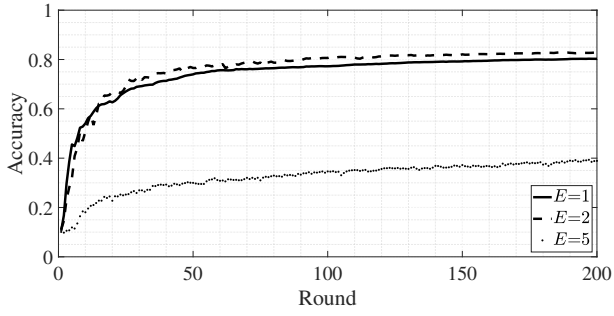


(d) Accuracy.

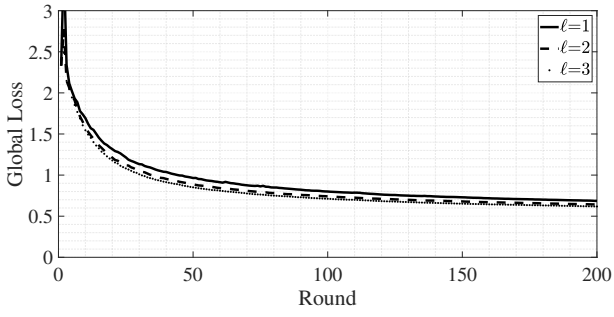
Fig. 4. Performance of the proposed FL with *Scheme I* sampling over MNIST dataset. In (a) and (b), $\ell = 1$. In (c) and (d), $E = 2$.



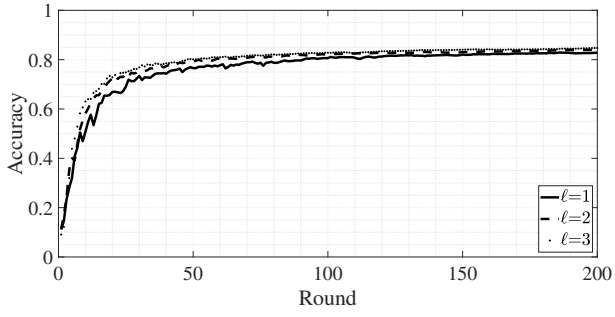
(a) Global loss.



(b) Accuracy.



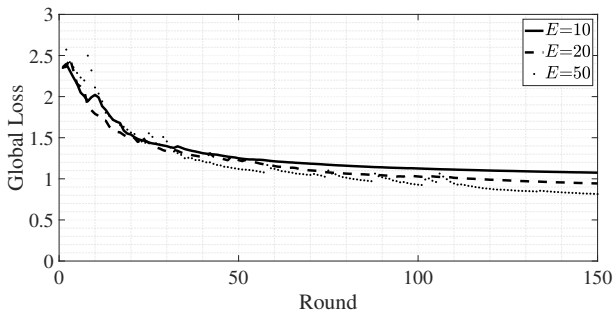
(c) Global loss.



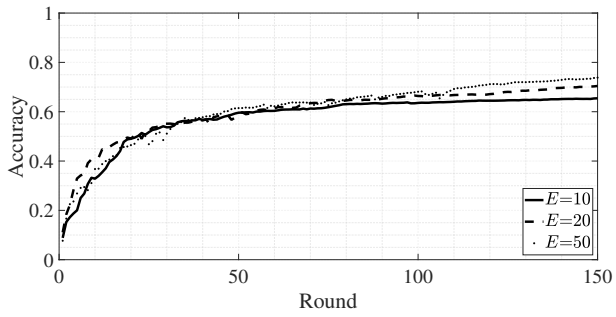
(d) Accuracy.

Fig. 5. Performance of the proposed FL with *Scheme II* sampling with $\hat{q}_k = \frac{1}{N}$, for all k, over MNIST dataset. In (a) and (b), $\ell = 1$. In (c) and (d), $E = 2$.

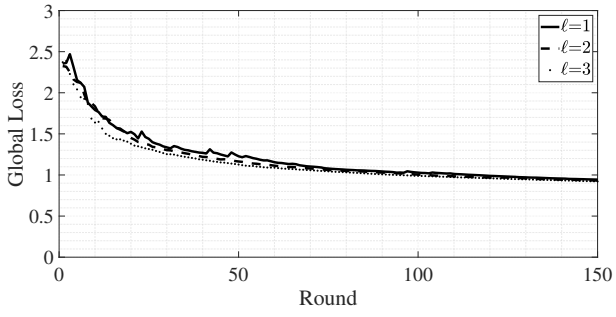
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications



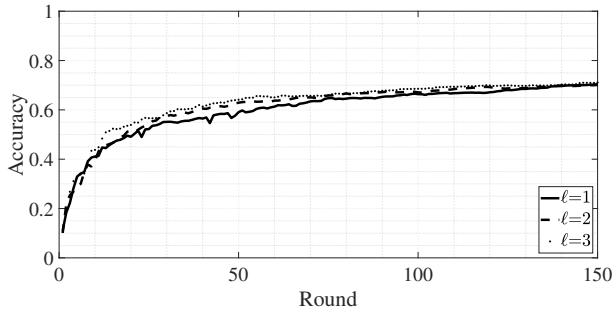
(a) Global loss.



(b) Accuracy.

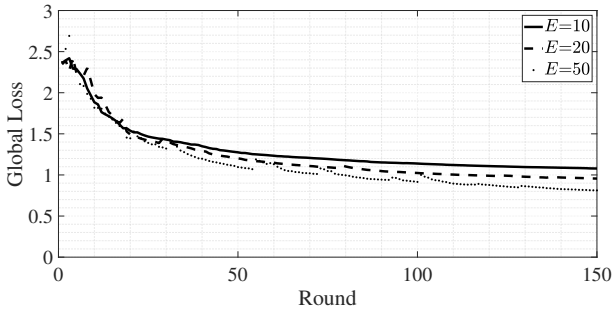


(c) Global loss.

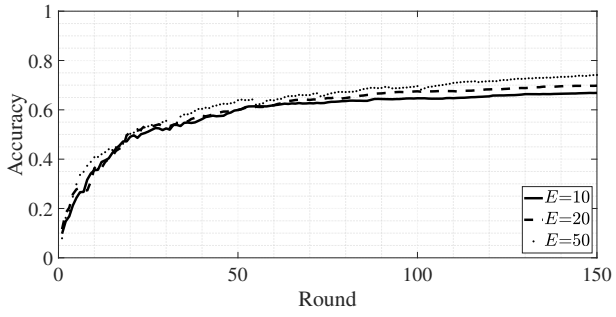


(d) Accuracy.

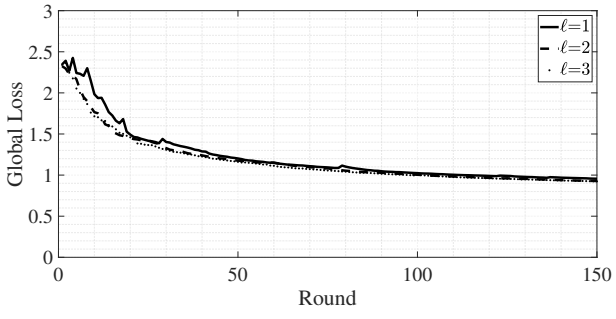
Fig. 6. Performance of the proposed FL with *Scheme I* sampling over the synthetic dataset. In (a) and (b), $\ell = 1$. In (c) and (d), $E = 20$.



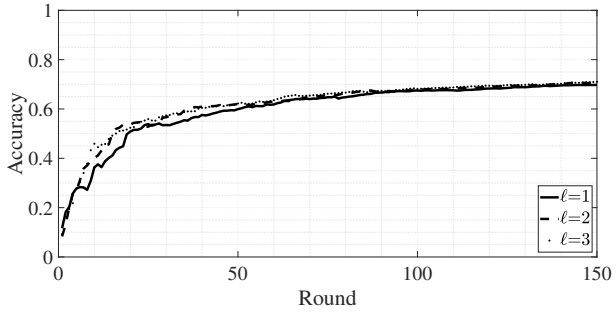
(a) Global loss.



(b) Accuracy.



(c) Global loss.



(d) Accuracy.

Fig. 7. Performance of the proposed FL with *Scheme II* sampling with $\hat{q}_k = \frac{1}{N}$, for all k, over the synthetic dataset. In (a) and (b), $\ell = 1$. In (c) and (d), $E = 20$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

wireless networks, where success probability varies across different devices. Our proposed federated learning algorithm incorporates the success probability in the averaging step. Thus, as the first step towards designing the FL algorithm, we have used stochastic geometry tools to calculate the success probability. We have studied the convergence of the proposed algorithm. Specifically, we have proven that the algorithm converges with rate $\mathcal{O}(\frac{1}{T})$ for strongly convex and smooth problems on non-i.i.d. data. The effects of computation, communication, and scheduling on the convergence rate of the algorithm have also been investigated. Finally, we have verified our algorithm through experimenting on real and synthetic datasets. FedAvg on non-i.i.d. data lacks theoretical guarantee in a convex optimization setting. Similarly, convergence analysis of our algorithm, which can be regarded as a modified version of FedAvg, with convex loss functions is an open problem.

Appendix A: Proof of Lemma 1

For device k at distance r_k from its associated BS, the success probability is derived as follows

$$
\begin{aligned}
U_k &= 1-\mathbb{E}\left[\mathbf{1}\left(\max\left\{\text{SINR}_k(1),...,\text{SINR}_k(\ell)\right\}<\theta\right)\mid k\in\mathcal{S}_t\right]\\
&= 1-\mathbb{E}_{\Phi_\text{I}}\left[\prod_{i=1}^{\ell}\mathbb{E}\left[\mathbf{1}\left(\text{SINR}_k(i)<\theta\right)\right]\mid k\in\mathcal{S}_t\right]\\
&= 1-\mathbb{E}_{\Phi_\text{I}}\left[\prod_{i=1}^{\ell}\mathbb{E}\left[\mathbf{1}\left(\frac{h_k(i)r_k^{-\alpha}}{I(i)+\sigma^2}<\theta\right)\right]\right]\\
&\overset{\text{(a)}}{=} 1-\mathbb{E}_{\Phi_\text{I}}\left[\left(1-e^{-\theta\sigma^2 r_k^\alpha}\prod_{x\in\Phi_\text{I}}\frac{1}{1+\theta r_k^\alpha\|x\|^{-\alpha}}\right)^{\ell}\right] \qquad \text{(A.1)}\\
&\overset{\text{(b)}}{=} \sum_{i=1}^{\ell}\binom{\ell}{i}(-1)^{i+1}e^{-i\theta\sigma^2 r_k^\alpha}\,\mathbb{E}_{\Phi_\text{I}}\left[\prod_{x\in\Phi_\text{I}}\frac{1}{(1+\theta r_k^\alpha\|x\|^{-\alpha})^i}\right]\\
&\overset{\text{(c)}}{\approx} \sum_{i=1}^{\ell}\binom{\ell}{i}(-1)^{i+1}\exp\Bigg\{-i\theta\sigma^2 r_k^\alpha\\
&\quad -\int_{\mathbb{R}^2}\left(1-\frac{1}{(1+\theta r_k^\alpha\|x\|^{-\alpha})^i}\right)\lambda(1-e^{-12/5\lambda\pi\|x\|^2})\mathrm{d}x\Bigg\}\\
&\overset{\text{(d)}}{=} \sum_{i=1}^{\ell}\binom{\ell}{i}(-1)^{i+1}\exp\Bigg\{-i\theta\sigma^2 r_k^\alpha-2\pi\lambda\\
&\quad\times\int_0^{\infty}\left(1-\frac{1}{(1+\theta r_k^\alpha x^{-\alpha})^i}\right)(1-e^{-12/5\lambda\pi x^2})x\mathrm{d}x\Bigg\},
\end{aligned}
$$

where (a) is obtained by averaging with respect to the fading. (b) follows from using the binomial expansion. In reality, Φ_I is a subset of the set of the participating devices in the federated learning. Since N is large in real-world applications, selecting a device in each cell out of a massive number of devices for assigning a resource block at the beginning of each FL round is almost identical to uniformly selecting a point inside the cell. This model is referred to as *model of type I* in [37]. By studying the pair correlation function (pcf) between a typical BS and interfering devices in this model, [37], [38] approximated Φ_I by a non-homogeneous PPP with intensity function $\lambda(x)=\lambda\left(1-e^{-12/5\lambda\pi\|x\|^2}\right)$. In (c), we use the probability generating functional (PGFL) for the point process Φ_I. Finally, (d) is obtained by using the polar domain representation.

Finally, note that averaging the success probability over r_k blurs the distinction between devices and removes the communication heterogeneity. In designing FL algorithms, however, we must incorporate the systems heterogeneity. Specifically, in FL algorithms, we must allocate higher weights to devices with lower success probabilities at the aggregation step which would not be possible if we averaged out r_k.

Appendix B: Proof of Theorem 1[9]

When we choose the learning rate as $\eta_t=\frac{\beta}{\mu(\gamma+t)}$ with $\beta>0$ and $\gamma=\max\left\{4\beta\frac{L}{\mu},E\right\}$, it fulfills the following properties: i) η_t is decreasing with respect to t, ii) $\eta_t\leq\frac{1}{2L}$, $\forall t$, and iii) $\eta_t\leq 2\eta_{t+E}$, $\forall t$. These properties help us with the convergence analysis of the proposed federated learning algorithm.

In the following, we provide three lemmas which help us in proving the theorem.

Lemma 2. *Averaging step in the proposed algorithm is unbiased, i.e.*

$$\mathbb{E}\left[\bar{w}_t\right]=\bar{v}_t,\qquad t\in\mathcal{I}_E,$$

where the expectation is over sampling and success event.

Proof: When $t\in\mathcal{I}_E$, $\bar{w}_t=w_t$; hence, from (8), we have

$$
\begin{aligned}
\mathbb{E}\left[\bar{w}_t\right] &= \mathbb{E}\Bigg[w_{t-E}+\sum_{k=1}^{N}\sum_{m=1}^{M}\frac{p_k}{q_k U_k}\\
&\quad\times\mathbf{1}\left(k\in\mathcal{S}_{t-E}(m),\text{SINR}_{k,m}>\theta\right)\left(v_t^k-w_{t-E}\right)\Bigg]\\
&\overset{\text{(a)}}{=} w_{t-E}+\sum_{k=1}^{N}\frac{p_k}{q_k U_k}\mathbb{E}\Bigg[\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}_{t-E}(m)\right)\\
&\quad\times\mathbb{E}\left[\mathbf{1}\left(\text{SINR}_{k,m}>\theta\right)\mid k\in\mathcal{S}_{t-E}(m)\right]\Bigg]\left(v_t^k-w_{t-E}\right)\\
&= w_{t-E}+\sum_{k=1}^{N}\frac{p_k}{q_k U_k}\\
&\quad\times\mathbb{E}\left[\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}_{t-E}(m)\right)U_k\right]\left(v_t^k-w_{t-E}\right)\\
&= w_{t-E}+\sum_{k=1}^{N}\frac{p_k}{q_k U_k}q_k U_k\left(v_t^k-w_{t-E}\right)\\
&= \sum_{k=1}^{N}p_k v_t^k=\bar{v}_t,
\end{aligned}
$$

where, in (a), the inner expectation is over the success event, and the outer expectation is with respect to the sampling. ■

Lemma 3. *When η_t satisfies the aforementioned properties,*

$$\mathbb{E}\left[\|\bar{w}_{t+1}-\bar{v}_{t+1}\|^2\right]\leq 4\eta_t^2E^2G^2B,\qquad t+1\in\mathcal{I}_E,$$

[9]We follow the same procedure as in [6]. Therefore, we mainly focus on the parts that are different.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

where $B=\sum_{k=1}^{N} p_k\left(\frac{1}{q_k U_k}-1\right)$ *for sampling Scheme I, and* $B=\sum_{k=1}^{N} p_k\left(\frac{1}{q_k U_k}-\frac{1}{M}\right)$ *for sampling Scheme II.*

Proof: For brevity, in this proof, we denote the set of scheduled devices at time $t+1-E$ by $\mathcal{S}$ rather than $\mathcal{S}_{t+1-E}$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{w}_{t+1}-\bar{v}_{t+1}\right\|^2\right]\\
&=\mathbb{E}\Bigg[\Bigg\|w_{t+1-E}+\sum_{k=1}^{N}\sum_{m=1}^{M}\frac{p_k}{q_k U_k}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\\
&\quad\times(v_{t+1}^k-w_{t+1-E})-\sum_{k=1}^{N}p_k v_{t+1}^k\Bigg\|^2\Bigg]\\
&=\mathbb{E}\Bigg[\Bigg\|\sum_{k=1}^{N}\sum_{m=1}^{M}\frac{p_k}{q_k U_k}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\\
&\quad\times(v_{t+1}^k-w_{t+1-E})-\sum_{k=1}^{N}p_k(v_{t+1}^k-w_{t+1-E})\Bigg\|^2\Bigg]\\
&=\mathbb{E}\Bigg[\Bigg\|\sum_{k=1}^{N}p_k\frac{\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)-q_k U_k}{q_k U_k}\\
&\quad\times(v_{t+1}^k-w_{t+1-E})\Bigg\|^2\Bigg]\\
&\overset{(a)}{\leq}\mathbb{E}\Bigg[\sum_{k=1}^{N}p_k\Bigg\|\frac{\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)-q_k U_k}{q_k U_k}\\
&\quad\times(v_{t+1}^k-w_{t+1-E})\Bigg\|^2\Bigg]\\
&\overset{(b)}{=}\sum_{k=1}^{N}p_k\frac{\mathbb{E}\left[\left(\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right)^2\right]-q_k^2U_k^2}{q_k^2U_k^2}\\
&\quad\times\mathbb{E}\left[\left\|v_{t+1}^k-w_{t+1-E}\right\|^2\right],
\end{aligned}\tag{B.1}$$

where (a) is obtained from convexity of $\|.\|^2$, and (b) follows from

$$\mathbb{E}\left[\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right]=q_k U_k,$$

where the expectation is with respect to the sampling and success event. In the following, we first calculate $\mathbb{E}\left[\left\|v_{t+1}^k-w_{t+1-E}\right\|^2\right]$, and then calculate $\mathbb{E}\left[\left(\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right)^2\right]$ for each sampling scheme separately.

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{t+1}^k-w_{t+1-E}\right\|^2\right]=\mathbb{E}\left[\left\|\sum_{i=t+1-E}^{t}\eta_i\nabla F_k(w_i^k;\xi_i^k)\right\|^2\right]\\
&\overset{(a)}{\leq}\mathbb{E}\left[E\sum_{i=t+1-E}^{t}\left\|\eta_i\nabla F_k(w_i^k;\xi_i^k)\right\|^2\right]\\
&\overset{(b)}{\leq}\mathbb{E}\left[\eta_{t+1-E}^2E\sum_{i=t+1-E}^{t}\left\|\nabla F_k(w_i^k;\xi_i^k)\right\|^2\right]\\
&\overset{(c)}{\leq}\eta_{t+1-E}^2E^2G^2\overset{(d)}{\leq}4\eta_{t+1}^2E^2G^2\leq 4\eta_t^2E^2G^2,
\end{aligned}\tag{B.2}$$

where (a) is obtained by using Cauchy-Schwarz inequality. (b) follows from the property that η_t is decreasing with respect to t. In (c), we have used **Assumption 4**. (d) follows from $\eta_t\leq 2\eta_{t+E}$.

At the aggregation step with *Scheme I* sampling, each device can use at most one resource block. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right)^2\right]\\
&=\mathbb{E}\left[\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right]=q_k U_k.
\end{aligned}\tag{B.3}$$

However, in the second scheme, the BS may allocate more than one resource block to a device at a sampling time. $\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)$ is a Bernoulli random variable which takes value one with probability $\hat{q}_k U_k$. For *Scheme II*, sampling and success event over each resource block are i.i.d.; therefore, $\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)$ is distributed according to a binomial distribution with parameters M and $\hat{q}_k U_k$, and we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{m=1}^{M}\mathbf{1}\left(k\in\mathcal{S}(m),\mathrm{SINR}_{k,m}>\theta\right)\right)^2\right]\\
&=q_k U_k\left(1-\frac{1}{M}q_k U_k\right)+q_k^2U_k^2\\
&=q_k U_k+\left(1-\frac{1}{M}\right)q_k^2U_k^2.
\end{aligned}\tag{B.4}$$

■

Lemma 4. *When* η_t *satisfies the aforementioned properties, for any* t*, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{v}_{t+1}-w^*\right\|^2\right]&\leq(1-\mu\eta_t)\mathbb{E}\left[\left\|\bar{w}_t-w^*\right\|^2\right]\\
&\quad+\eta_t^2\left(\sum_{k=1}^{N}p_k^2\sigma_k^2+6L\Gamma+8(E-1)^2G^2\right).
\end{aligned}$$

Proof: See **Appendix A** in [6]. ■

Note that **Lemma 2** and **Lemma 3** only study the averaging steps ($t\in\mathcal{I}_E$), while **Lemma 4** is a result of one step of SGD (for any t). **Lemma 2** shows that each averaging step is unbiased, and **Lemma 3** shows that variance of each averaging step is bounded. When, $t\notin\mathcal{I}_E$, $\bar{w}_t=\bar{v}_t$ according to (8).

Based on the above discussion, at any t, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{w}_{t+1}-w^*\right\|^2\right]=\mathbb{E}\left[\left\|\bar{w}_{t+1}-\bar{v}_{t+1}+\bar{v}_{t+1}-w^*\right\|^2\right]\\
&\overset{(a)}{=}\mathbb{E}\left[\left\|\bar{w}_{t+1}-\bar{v}_{t+1}\right\|^2\right]+\mathbb{E}\left[\left\|\bar{v}_{t+1}-w^*\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left(\bar{w}_{t+1}-\bar{v}_{t+1}\right)^T\left(\bar{v}_{t+1}-w^*\right)\right]\\
&\overset{(b)}{\leq}(1-\mu\eta_t)\mathbb{E}\left[\left\|\bar{w}_t-w^*\right\|^2\right]+\eta_t^2\Bigg(\sum_{k=1}^{N}p_k^2\sigma_k^2+6L\Gamma\\
&\quad+8(E-1)^2G^2+4E^2G^2B\Bigg),
\end{aligned}\tag{B.5}$$

where the last term in (a) is zero based on **Lemma 2** when $t\in\mathcal{I}_E$ and the fact that $\bar{w}_t=\bar{v}_t$ when $t\notin\mathcal{I}_E$. In (b), we have used **Lemma 3** and **Lemma 4**.

For brevity, we define $C = \sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2 + 4E^2 G^2 B$ and $\Delta_t = \|\bar{w}_t - w^*\|^2$. In the following, we find v such that $\mathbb{E}[\Delta_t] \leq \frac{v}{\gamma+t}$ at any t after initializing with Δ_0. This is satisfied at time $t = 0$ when $v \geq \gamma \Delta_0$. Moreover, when $\beta > 1$ and $v \geq \frac{\beta^2 C}{\mu^2(\beta-1)}$, $\mathbb{E}[\Delta_{t+1}] \leq \frac{v}{\gamma+t+1}$ given $\mathbb{E}[\Delta_t] \leq \frac{v}{\gamma+t}$. The proof is as follows:

$$\begin{aligned}
\mathbb{E}[\Delta_{t+1}] &\leq (1-\mu\eta_t)\mathbb{E}[\Delta_t] + \eta_t^2 C \\
&\leq \left(1 - \frac{\beta}{\gamma+t}\right)\frac{v}{\gamma+t} + \frac{\beta^2 C}{\mu^2(\gamma+t)^2} \\
&= \frac{\gamma+t-1}{(\gamma+t)^2} v + \left[\frac{\beta^2 C}{\mu^2(\gamma+t)^2} - \frac{\beta-1}{(\gamma+t)^2} v\right] \\
&\overset{(a)}{\leq} \frac{\gamma+t-1}{(\gamma+t)^2} v \leq \frac{v}{\gamma+t+1},
\end{aligned}$$

where (a) is obtained from $v \geq \frac{\beta^2 C}{\mu^2(\beta-1)}$. Thus, by induction $\mathbb{E}[\Delta_t] \leq \frac{v}{\gamma+t}$ at any t when $v = \max\left\{\frac{\beta^2 C}{\mu^2(\beta-1)}, \gamma\Delta_0\right\}$.

Therefore, when $\eta_t = \frac{2}{\mu(\gamma+t)}$[10] with $\gamma = \max\{8\frac{L}{\mu}, E\}$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{w}_t - w^*\|^2\right] \leq \frac{1}{\gamma+t} \max\Bigg\{ &\frac{4}{\mu^2}\Bigg(\sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma \\
&+ 8(E-1)^2 G^2 + 4E^2 G^2 B\Bigg), \gamma \|w_0 - w^*\|^2 \Bigg\}, \quad \text{(B.6)}
\end{aligned}$$

where $B = \sum_{k=1}^{N} p_k \left(\frac{1}{q_k U_k} - 1\right)$ for sampling *Scheme I*, and $B = \sum_{k=1}^{N} p_k \left(\frac{1}{q_k U_k} - \frac{1}{M}\right)$ for sampling *Scheme II*.

After the averaging step at time T ($T \in \mathcal{I}_E$), from L-smoothness of the global objective function F, we have

$$\begin{aligned}
\mathbb{E}[F(w_T) - F^*] &\leq \frac{L}{2}\mathbb{E}\left[\|w_T - w^*\|^2\right] \\
&\overset{(a)}{\leq} \frac{L}{2(\gamma+T)} \max\Bigg\{ \frac{4}{\mu^2}\Bigg(\sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2 \\
&\quad + 4E^2 G^2 B\Bigg), \gamma \|w_0 - w^*\|^2 \Bigg\} \\
&\leq \frac{L/\mu}{\gamma+T}\Bigg[\frac{2}{\mu}\Bigg(\sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2 \\
&\quad + 4E^2 G^2 B\Bigg) + \frac{\mu\gamma}{2}\|w_0 - w^*\|^2\Bigg],
\end{aligned}$$

where (a) is obtained from $w_T = \bar{w}_T$ and (B.6).

Appendix C

With slight abuse of notation, we define $\hat{F}(w) = \sum_{k=1}^{N} \frac{p_k U_k}{\sum_{k'=1}^{N} p_{k'} U_{k'}} F_k(w)$, and we prove that (18) solves

$$\min_{w} \quad \hat{F}(w) = \sum_{k=1}^{N} \frac{p_k U_k}{\sum_{k'=1}^{N} p_{k'} U_{k'}} F_k(w), \quad \text{(C.1)}$$

which has the same solution as (19). Let us denote the solution to (C.1) by $\hat{w}^*$, and define $\alpha_k = p_k U_k$, $\alpha = \sum_{k=1}^{N} \alpha_k$, and $\alpha'_k = \frac{\alpha_k}{\alpha}$; therefore, $\sum_{k=1}^{N} \alpha'_k = 1$. For brevity, we also define $H_{k,m} = \mathbf{1}\left(k \in \mathcal{S}(m), \text{SINR}_{k,m} > \theta\right)$, where we have ignored the time index of $\mathcal{S}$ since it is i.i.d. over different sampling steps.

Since we are using a different averaging approach, we must check **Lemma 2** and **Lemma 3**. However, **Lemma 4** is not affected by the averaging steps; thus, it still holds (after replacing p_k with α'_k). We also need to change p_k to α'_k in definitions of $\bar{v}_t$, $\bar{w}_t$, g_t, and $\bar{g}_t$. It is worth reminding that, for sampling, we use *Scheme II* with $\hat{q}_k = p_k$.

Moreover, according to (18), when there is no successful transmission, the global model parameters at the BS do not change, which only affects the convergence rate (not the converging point). Therefore, to prove the convergence of (18) to $\hat{w}^*$, we assume $\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0$, i.e. at least one local update is available at the BS at each averaging step. Given $\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0$, we can write (18) as

$$w_t = \sum_{k=1}^{N}\sum_{m=1}^{M} \frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}} v_t^k, \qquad t \in \mathcal{I}_E.$$

In the following, we provide a lemma that helps us derive **Lemma 2** and **Lemma 3** for the new averaging approach.

Lemma 5. *For Scheme II sampling with $\{\hat{q}_k = p_k\}$, we have*

$$\mathbb{E}\left[\sum_{m=1}^{M} \frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}} \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0\right] = \alpha'_k.$$

Proof:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{m=1}^{M} \frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}} \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0\right] \\
&= \sum_{m=1}^{M} \mathbb{E}_{H_{k,m}}\Bigg[\mathbb{E}\Bigg[\frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}} \\
&\qquad \Bigg| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0, H_{k,m}\Bigg]\Bigg] \\
&= \sum_{m=1}^{M} \mathbb{P}\left(H_{k,m} = 1 \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0\right) \\
&\quad \times \mathbb{E}\Bigg[\frac{1}{1 + \sum_{k'=1}^{N}\sum_{\substack{m'=1, \\ m' \neq m}}^{M} H_{k',m'}} \Bigg| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0, \\
&\qquad\qquad H_{k,m} = 1\Bigg] \\
&= \sum_{m=1}^{M}\sum_{i=0}^{M-1} \frac{1}{1+i} \mathbb{P}\Bigg(\sum_{k'=1}^{N}\sum_{\substack{m'=1, \\ m' \neq m}}^{M} H_{k',m'} = i, H_{k,m} = 1 \\
&\qquad\qquad \Bigg| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0\Bigg). \quad \text{(C.2)}
\end{aligned}$$

When $H_{k,m} = 1$, we have $\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'} > 0$. Moreover, resource allocation is i.i.d. over difference resource

[10] We set $\beta = 2$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3081746, IEEE Transactions on Communications

blocks, i.e. random variable $\sum_{k'=1}^{N}\sum_{\substack{m'=1,\\ m'\neq m}}^{M} H_{k',m'}$ is independent of random variable $H_{k,m}$. Thus,

$$\mathbb{P}\left(\sum_{k'=1}^{N}\sum_{\substack{m'=1,\\ m'\neq m}}^{M} H_{k',m'}=i, H_{k,m}=1 \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right)$$
$$=\frac{\mathbb{P}\left(H_{k,m}=1\right)}{\mathbb{P}\left(\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right)} \times\mathbb{P}\left(\sum_{k'=1}^{N}\sum_{\substack{m'=1,\\ m'\neq m}}^{M} H_{k',m'}=i\right)$$
$$\overset{\text{(a)}}{=}\frac{\alpha_k}{1-(1-\alpha)^M}\binom{M-1}{i}\alpha^i(1-\alpha)^{M-1-i}, \tag{C.3}$$

where (a) is obtained using $\mathbb{P}\left(H_{k,m}=1\right)=\alpha_k$ and $\mathbb{P}\left(\sum_{k=1}^{N} H_{k,m}=1\right)=\alpha$. Finally, **Lemma 5** is obtained by substituting (C.3) in (C.2). ∎

When $t\in\mathcal{I}_E$, we have $\bar{w}_t=w_t$; thus,

$$\mathbb{E}\left[\bar{w}_t \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right]=$$
$$\mathbb{E}\left[\sum_{k=1}^{N}\sum_{m=1}^{M}\frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}}v_t^k \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right]$$
$$=\sum_{k=1}^{N}\mathbb{E}\left[\sum_{m=1}^{M}\frac{H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}} \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right] \times v_t^k$$
$$\overset{\text{(a)}}{=}\sum_{k=1}^{N}\alpha'_k v_t^k=\bar{v}_t, \tag{C.4}$$

where the expectation is with respect to sampling and success event and (a) follows from **Lemma 5**.

At $t+1\in\mathcal{I}_E$, we also have

$$\mathbb{E}\left[\left\|\bar{w}_{t+1}-\bar{v}_{t+1}\right\|^2 \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right]$$
$$=\mathbb{E}\left[\left\|\sum_{k=1}^{N}\frac{\sum_{m=1}^{M} H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}}\left(v_{t+1}^k-\bar{v}_{t+1}\right)\right\|^2 \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right]$$
$$\leq\mathbb{E}\left[\sum_{k=1}^{N}\frac{\sum_{m=1}^{M} H_{k,m}}{\sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}}\left\|v_{t+1}^k-\bar{v}_{t+1}\right\|^2 \middle| \sum_{k'=1}^{N}\sum_{m'=1}^{M} H_{k',m'}>0\right]$$
$$\overset{\text{(a)}}{=}\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left\|v_{t+1}^k-\bar{v}_{t+1}\right\|^2\right]$$
$$=\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left\|\left(v_{t+1}^k-w_{t+1-E}\right)-\left(\bar{v}_{t+1}-w_{t+1-E}\right)\right\|^2\right]$$
$$=\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left(\left\|v_{t+1}^k-w_{t+1-E}\right\|^2+\left\|\bar{v}_{t+1}-w_{t+1-E}\right\|^2 -2\left(v_{t+1}^k-w_{t+1-E}\right)^T\left(\bar{v}_{t+1}-w_{t+1-E}\right)\right)\right]$$
$$=\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left\|v_{t+1}^k-w_{t+1-E}\right\|^2\right]-\mathbb{E}\left[\left\|\bar{v}_{t+1}-w_{t+1-E}\right\|^2\right]$$
$$\leq\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left\|v_{t+1}^k-w_{t+1-E}\right\|^2\right]$$
$$=\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k\left\|\sum_{i=t+1-E}^{t}\eta_i\nabla F_k(w_i^k,\xi_i^k)\right\|^2\right]$$
$$\leq\mathbb{E}\left[\sum_{k=1}^{N}\alpha'_k E\sum_{i=t+1-E}^{t}\left\|\eta_i\nabla F_k(w_i^k,\xi_i^k)\right\|^2\right]$$
$$\leq\sum_{k=1}^{N}\alpha'_k E^2\eta_{t+1-E}^2 G^2\leq 4\eta_t^2E^2G^2, \tag{C.5}$$

where, to drive (a), we use **Lemma 5**. The last line is also obtained similar to (B.2).

From (C.4), we understand that **Lemma 2** holds (for problem (C.1)). Also, from (C.5), we understand that **Lemma 3** holds with $B=1$. As we discussed earlier, **Lemma 4** is also valid (after replacing p_k with α'_k). Thus, following the same procedure as in **Appendix B**, we can prove that with using (18) at averaging steps, at time $T\in\mathcal{I}_E$, we have

$$\mathbb{E}\left[\left\|w_T-\hat{w}^*\right\|^2\right]\leq\frac{1}{\gamma+T}\max\left\{\frac{4}{\mu^2}\left(\sum_{k=1}^{N}\alpha_k'^2\sigma_k^2+6L\Gamma +8(E-1)^2G^2+4E^2G^2\right),\gamma\left\|w_0-\hat{w}^*\right\|^2\right\}.$$

Hence, the algorithm converges to the solution to (C.1).

REFERENCES

[1] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[3] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *arXiv preprint arXiv:2009.13012*, 2020.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[5] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[6] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[7] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

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

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

[10] C. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *arXiv preprint arXiv:1910.13067*, 2019.

[11] F. Zhou and G. Cong, "On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization," *arXiv preprint arXiv:1708.01012*, 2017.

[12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[14] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," *arXiv preprint arXiv:1910.14425*, 2019.

[15] H. T. Nguyen, V. Sehwag, S. Hosseinalipour, C. G. Brinton, M. Chiang, and H. V. Poor, "Fast-convergent federated learning," *IEEE Journal on Selected Areas in Communications*, 2020.

[16] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 317–333, 2019.

[17] V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi, "Cocoa: A general framework for communication-efficient distributed optimization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8590–8638, 2017.

[18] J. G. Andrews, A. K. Gupta, and H. S. Dhillon, "A primer on cellular network analysis using stochastic geometry," *arXiv preprint arXiv:1604.03183*, 2016.

[19] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.

[20] H. ElSawy, A. Sultan-Salem, M.-S. Alouini, and M. Z. Win, "Modeling and analysis of cellular networks using stochastic geometry: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 167–203, 2016.

[21] R. K. Ganti and M. Haenggi, "Asymptotics and approximation of the sir distribution in general cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2130–2143, 2015.

[22] C. Saha, M. Afshang, and H. S. Dhillon, "3gpp-inspired hetnet model using poisson cluster process: Sum-product functionals and downlink coverage," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2219–2234, 2017.

[23] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge University Press, 2012.

[24] R. Tanbourgi, H. S. Dhillon, J. G. Andrews, and F. K. Jondral, "Effect of spatial interference correlation on the performance of maximum ratio combining," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3307–3316, 2014.

[25] H. ElSawy and E. Hossain, "On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4454–4469, 2014.

[26] M. Di Renzo and P. Guan, "Stochastic geometry modeling and system-level analysis of uplink heterogeneous cellular networks with multi-antenna base stations," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2453–2476, 2016.

[27] T. Bai and R. W. Heath, "Analyzing uplink sinr and rate in massive mimo systems using stochastic geometry," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4592–4606, 2016.

[28] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, 2015.

[29] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3321–3363, 2013.

[30] S. U. Stich, "Local sgd converges fast and communicates little," *arXiv preprint arXiv:1805.09767*, 2018.

[31] D. S. Mitrinovic and P. M. Vasic, *Analytic Inequalities*. Berlin, Germany: Springer, 1970, vol. 61.

[32] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[33] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.1," 2014.

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *International Conference on Machine Learning*, 2014, pp. 1000–1008.

[36] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," https://github.com/mhdslh/Federated-Learning-in-Unreliable-and-Resource-Constrained-Cellular-Wireless-Networks, 2020.

[37] M. Haenggi, "User point processes in cellular networks," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 258–261, 2017.

[38] Y. Wang, M. Haenggi, and Z. Tan, "The meta distribution of the sir for cellular networks with power control," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1745–1757, 2017.

Mohammad Salehi received the B.Sc. degree in electrical engineering from K. N. Toosi University of Technology, Tehran, Iran, in 2014 and the M.Sc. degree in electrical engineering from Amirkabir University of Technology, Tehran, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Manitoba, Winnipeg, Canada. His research interests include modeling and analyzing wireless networks.

Ekram Hossain (F'15) Professor in the Department of Electrical and Computer Engineering at University of Manitoba, Canada (http://home.cc.umanitoba.ca/∼hossaina). He is a Member (Class of 2016) of the College of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Engineering Institute of Canada. Dr. Hossain's current research interests include design, analysis, and optimization of wireless networks with emphasis on beyond 5G cellular networks. He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks". He received the 2017 IEEE ComSoc TCGCC (Technical Committee on Green Communications & Computing) Distinguished Technical Achievement Recognition Award "for outstanding technical leadership and achievement in green wireless communications and networking". He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2017, 2018, 2019, and 2020. Currently he serves as the Editor-in-Chief of IEEE Press. Previously he served as the Editor-in-Chief for the IEEE Communications Surveys and Tutorials (2012–2016).