Rate-Convergence Tradeoff of Federated Learning Over Wireless Channels

Ayoob Salari, Graduate Student Member, IEEE, Sarah J. Johnson, Senior Member, IEEE, Branka Vucetic, Life Fellow, IEEE, and Mahyar Shirvanimoghaddam, Senior Member, IEEE

Abstract—In this article, we consider a federated learning (FL) problem over wireless channel that takes into account the coding rate and packet transmission errors. Communication channels are modeled as packet erasure channels (PECs), where the probability of erasure is determined by block length, code rate, and signal-to-noise ratio (SNR). In spite of fluctuations in instantaneous loss of FL, we prove that the expectation of loss converges even in the presence of packet erasure. To mitigate the impact of packet erasure on FL performance, we suggest a paradigm in which the central node (CN) makes use of memory. In particular, we propose two schemes in which, in the event of packet erasure, the CN retains either the most recent local updates or some of packet erasure on FL performance, we examine a realistic scenario of a massive IoT under the assumption of error-prone transmissions. Our simulation results demonstrate that even a single memory unit has a considerable effect on the FL’s efficiency in erroneous communication.

Index Terms—Channel coding, convergence, federated learning (FL), massive Internet of Things (mIoT), packet erasure, uplink.

I. INTRODUCTION

INTERNET of Things (IoT)-enabled applications and services have grown in popularity due to their potential to improve human lives [1]. By the year 2023, there will be 14.7 billion connected devices, with IoT devices making up half of that total [2]. It is envisioned that the number of concurrent connections increase from one million per square kilometer in 5G to ten million per square kilometer in 6G [3], [4]. Together with artificial intelligence (AI), the next generation of wireless technologies intends to enable the communication infrastructure required for the massive IoT (mIoT) use cases [5], [6]. In mIoT, a large number of low-cost low-power endpoints communicate with the central node (CN) [7]. Conventionally, users send their local data to a computationally capable central controller for the purpose of training a deep network model. However, such an approach is impractical in an mIoT scenario owing to concerns about privacy, limited power, and often inadequate communication bandwidth, all of which impose a significant strain on the communication links [8]. Federated learning (FL) is an alternative approach, in which each user trains a local model based on its own data and the global parameter and transmits the updated model parameter to the CN. The CN adjusts the global model parameter with a weighted average of users’ updates and broadcasts the new global parameters to the network. This procedure is continued until convergence occurs [9].

It is critical to comprehend the difficulties associated with FL. Since each device’s data set is acquired by the device itself, it depends on the client’s local environment. As a result, not only the users’ data sets are non-i.i.d throughout the network but also the size of users’ data sets may vary significantly [10]. This statistical heterogeneity has an impact on the convergence mechanism of FL and lowers model accuracy. One can use an adaptive averaging strategy or apply a data-sharing mechanism to reduce the impact of non-i.i.d data sets [11]. To cope with the heterogeneity of systems, FL must handle a variety of devices that have differing amounts of memory and processing power as well as differing battery sizes and storage capacities [12]. Weight-based federated averaging is considered as a solution to this problem [13]. Another challenge of FL is the tradeoff between a devices processing power and the communication overhead. Local processing is substantially faster than communication in the network. Also, increasing the number of local computing iterations leads to a decrease in the number of network communications. On the other hand, since users are constrained in terms of power, increasing the number of local iterations depletes the battery and the device’s ability to communicate with the CN is severely limited [14].

While substantial effort has been made to tackle the aforementioned FL challenges, it is typically assumed that one could simply use FL in wireless networks and devices could communicate with the CN without error [10], [11], [12], [13], [14], [15]. However, in an mIoT configuration, where a large number of power- and bandwidth-limited devices with low-computing capacity transmit short-packet data to a CN, the wireless channels are unreliable and communication is often erroneous [16]. Despite the recent investigations of the FL algorithms over wireless fading channels, they are constrained in a number of ways [17], [18], [19]. The study in [18] only indicated a drop in model accuracy in the event of...
a communication failure but did not suggest any strategies for improvement. To alleviate the impact of fading channels, Amiri and Gündüz [17] considered only one device is transmitting in each iteration; however, in an mIoT scenario where a large number of users desire to connect with the CN, this strategy is unsuitable. Chen et al. [19] presented resource allocation for minimizing model loss in case of communication error. However, they neglected the network’s statistical heterogeneity, which has a significant effect on the system performance. Furthermore, the accuracy of their suggested model for handwritten digit identification is 75%. Shirvanimoghaddam et al. [20] investigated the FL algorithm over packet erasure channels (PECs) and demonstrated that when the CN depends solely on fresh local updates, both the loss and accuracy of the model fluctuate. However, it is yet unknown how the physical-layer parameters, such as the code rate and blocklength, affect the convergence of FL in erroneous communication.

In this article, we consider a realistic scenario of FL in mIoT that takes both the coding rate and the packet transmission error into consideration. We model the communication channel as a PEC, in which the CN either successfully receives local updates from devices or, with a certain probability, the packet is erased. We first assume the CN has no memory and demonstrate that although FL’s performance converges on average, it varies from instant to instant. To mitigate the impact of erasures on FL convergence, we assume that the CN stores the model parameters in memory. We investigate two different schemes: 1) CN has sufficient memory to record the last local update of every device and 2) CN has limited capacity and caches the previous m global updates. We analyze two distinct communication scenarios, i.e., short-packet and long-packet communications. The erasure probability is determined by the blocklength, code rate, and signal-to-noise ratio (SNR). We show that, for fixed transmission power, lowering the code rate reduces the probability of error while increasing the reliability. However, because of the low code rate, there will be more communications between the CN and the devices. Therefore, there is a tradeoff between code rate, convergence time, and the FL model accuracy.

Our main contributions in this article are summarized as follows.

1) We analyze a real-world FL scenario in which communication is erroneous. In our model, the CN either successfully receives local updates from the devices or, with a certain probability, the packet is wiped along the way. We will represent the global updates formula for two unique cases: 1) CN with memory and 2) CN without memory. For both FL cases, we apply the saddle-point approximation to determine the probability distribution of the global update in each communication round. In addition, we will prove that the expected total loss of the FL will converge in the case that the CN does not have memory. Furthermore, we prove that FL converges to the global minimum of the loss function when the CN possesses memory to store past local or global updates.

2) We examine two distinct approaches in the scenario where the CN has memory. One solution involves CN remembering each user’s prior local update, while another involves CN storing global updates. We demonstrate the impact of blocklength, coding rate, and SNR on the convergence of FL for both approaches. We illustrate that having even one memory unit at the CN has a significant impact on the performance of the system. We also show that increasing the number of communication rounds between the CN and the users or decreasing the coding rate does not necessarily result in the minimizing the loss function for a given training duration. In reality, the coding rate should be carefully chosen to minimize loss for each SNR regime.

3) We investigate the performance of FL in erroneous communication for two practical scenarios. First, we examine an mIoT configuration based on actual measurements [21], in which 196 IoT devices perform measurements and use short packets to communicate with the CN. We calculate the packet error rate using the normal approximation as a function of the block-length, coding rate, and SNR [22] and demonstrate the required time for the convergence of the FL in such a scenario based on rate and SNR. We then evaluate a scenario involving long-packet communications, in which IoT devices use FL to train neural networks using the MNIST digits data set [23]. In this scenario, we check to determine whether the instantaneous received SNR is greater than the SNR threshold necessary for error-free transmission and demonstrate the effect of rate and SNR on model accuracy.

The remainder of this article is structured as follows. Section II describes the system model and the various FL techniques used in erroneous communications. Section III analyzes the performance of the FL algorithm with erroneous communications. Section IV contains numerical results. Section V presents a real-world FL scenario and the performance of the proposed approaches. Finally, Section VI concludes this article.

Notations: The operators E[·], ∇, and [·] indicate the expectation, the gradient, and the sequence product, respectively. The notations |a|2 denote the ℓ2-norm of a.

II. SYSTEM MODEL

We consider an FL system comprising of one CN and a set U of U IoT devices with local data sets, D1, D2, , Dm. The data set of each device u is defined as Dv = (xu1, yu1, xu2, yu2, , xu, yu), where Du is the size of the user u’s data set. The total amount of training data stored by all users is stated as D = i Du. The loss function, which varies depending on the learning model, is used to evaluate the FL algorithm’s performance. For a linear regression learning model, the loss function can be written as f(ω, x, y) = (1/2)∥y − ωTx∥2, while in the case of neural network, it is f(ω, x, y) = (1/2)∥y − fn(x; ω)∥2, where fn(x; ω) is the learning output of the neural network and ω is the weight vector to be learned.

Suppose IoT devices wish to send their local updates in messages with a length of k bits to the CN. An encoder is used to map these messages to codewords of length n using a...
channel code of rate $R = k/n$. The received signal of user $u$ at the CN is
\[ r_u = h_u z_u + n \] (1)
where $r_u$ is the received signal, $z_u$ is the transmitted signal of user $u$, and $n$ is zero-mean additive white Gaussian noise (AWGN) with variance $\sigma^2$ and $h_u$ is the channel gain that follows a zero-mean circularly symmetric Gaussian distribution, i.e., $h_u \sim \text{CN}(0, 1)$. We presume that all packets experience independently and identically distributed Rayleigh fading. We further assume the channel is block fading, which is constant over each packet duration and independently varies across packets. For short packets, particularly in mMTC applications, the receiver can be written as
\[ \omega \]
short- and long-packet communication. In the case of long-packet transmission, the packet error rate can be accurately estimated by using the outage probability as follows:
\[ \epsilon(\gamma, R) = \Pr(\gamma < \gamma_{th}) \]
(2)
where $\gamma_{th} = 2^R - 1$. For the short-packet communication, Polyanskiy et al. [22] showed that the packet error rate at the receiver can be written as
\[ \epsilon(\gamma, n, R) \approx Q\left(\frac{nC(\gamma) - k + 0.5\log_2(n)}{\sqrt{nV(\gamma)}}\right) \]
(3)
where $C(\gamma) = \log_2(1 + \gamma)$ is the channel capacity, $V(\gamma) = \log_2^2(\epsilon)(1 - (1 + \gamma)^{-2})$ is the channel dispersion, and $Q(.)$ is the standard Q-function [22].

While uplink transmissions are erroneous, we assume the CN utilizes the whole spectrum and transmits with high power on the downlink; therefore, the error in the downlink direction is negligible.

A. Principles of FL in Error-Free Scenario

In each iteration of FL, each device $u$ computes its local update $\omega_u$ (usually using a few epochs of gradient descent (GD)), and transmits its updated parameter of the trained model to the CN. Next, the CN calculates the average weight, $\omega$, by aggregating all the local updates. Then, the CN broadcasts the global parameter to be used by devices for the next iteration of FL.

Let us consider the loss function of device $u$, which calculates the model error on its data set $D_u$, as
\[ F_u(\omega) = \frac{1}{D_u} \sum_{i=1}^{D_u} f(\omega, x_u^i, y_u^i), \]
(4)
Employing the GD approach, the local parameter of device $u$ at time $t$ can be computed as
\[ \omega_u^{(t)} = \omega^{(t-1)} - \eta \nabla F_u(\omega^{(t-1)}) \]
(5)
where $\eta$ is the learning rate. Once the IoT device has computed its own local parameter, it will transmit the updated parameter to the CN through an error-free channel, and the CN will aggregate all of the received local parameters to compute the global update using
\[ \omega^{(t)} = \frac{1}{D} \sum_{u=1}^{U} D_u \omega_u^{(t)} \]
(6)
After calculating the global parameter, the CN will broadcast it throughout the network. One could combine the last two steps of FL and calculate the global update as
\[ \omega^{(t)} = \omega^{(t-1)} - \frac{\eta}{D} \sum_{u=1}^{U} D_u \nabla F_u(\omega^{(t-1)}). \]
(7)

B. FL in the Presence of Communication Errors

Given the SNR, $\gamma$, blocklength, $n$, and code rate, $R$, the probability that the CN does not receive the local parameters can be calculated. We will utilize this to formulate the FL updating rules in the presence of communication errors.

1) FL With Erroneous Communications Without the CN Memory: The number of local parameters received by the CN during each communication round may vary depending on the channel quality. In general, when the CN does not have a memory to store past local/global updates, the global parameter at the $(t + 1)$th iteration can be calculated as
\[ \omega^{(t+1)} = \frac{\sum_{u=1}^{U} I_u^{(t)} D_u \omega_u^{(t)}}{\sum_{u=1}^{U} I_u^{(t)} D_u} \]
(8)
where $I_u^{(t)} \in [0, 1]$ indicates whether the CN receives the local update of user $u$ at time $t$ correctly, which follows a Bernoulli distribution:
\[ I_u^{(t)} = \begin{cases} 1 & \text{with probability } 1 - \epsilon(\gamma_u, n, R) \\ 0 & \text{with probability } \epsilon(\gamma_u, n, R). \end{cases} \]
In this scenario, the number of local updates at the CN may vary across iterations.$^1$

2) FL With Erroneous Communications With the CN Memory: We will assume in this part that the CN has a memory to store model parameters.

a) CN caches each user’s past local parameter: In this setup, we suppose the CN has a memory dedicated to storing each user’s most recent local parameter. To compute the global parameter at each communication round, the CN employs the fresh local parameter for users with successful transmissions and reuses the stored local parameters for users with an erroneous channel, i.e., in the case of channel erasure, $\omega_u^{(t)} = \omega_u^{(t-1)}$. Therefore, the number of users that participate in the update of global parameter would be fixed. The global update can be computed as
\[ \omega^{(t+1)} = \frac{1}{D} \sum_{u=1}^{U} D_u \left( \omega_u^{(t)} + \omega_u^{(t-1)} (1 - I_u^{(t)}) \right). \]
(9)

$^1$Partial recovery can be considered as an extension to the model. In such a case, the update rule would need to be altered. However, in the current model of this article, we consider that if we are unable to recover the original $k$ bits, we declare an error in the system.
updates, and this case, the global update at the time instant $t$

The global update can be written as $\omega(t) = \frac{1}{D} \sum_{u=1}^{U} D_u \left( \omega_u^{(t)} + (1 - f_u^{(t)}) \sum_{i=1}^{m} \alpha_{t-i} \omega^{(t-i)} \right)$.

where $m$ is the weighted average of the past $m$ global updates, and $\alpha_{t-i}$ represents the weight of the global parameter at time instant $t - i$. We choose $m = 1$.

Fig. 1 compares FL approaches in the presence of communication errors, where we assumed that $\epsilon_u = \epsilon$ for all users. It can be seen that the best performance is for the case where the CN has memory to store the past global parameters. When the CN has a restricted memory space, instead of saving all local parameters of IoT devices, the CN may store the previous $m$ global updates (Section II-B2b), however, the performance deteriorates in comparison to the scenario where the CN memorizes local values.

III. PERFORMANCE ANALYSIS

A. FL in Erroneous Communication Without the CN Memory

For the case with erroneous communications, where the CN does not have memory, the global update can be written as $\omega(t) = \frac{1}{D} \sum_{u=1}^{U} D_u \left( \omega_u^{(t)} + (1 - f_u^{(t)}) \sum_{i=1}^{m} \alpha_{t-i} \omega^{(t-i)} \right)$.

where, we assume that all devices have the same training data set size, i.e., $D_u = D/U$. Let us consider the numerator of (11) as $Y = \sum_{u=1}^{U} t_u^{(t)} \omega_u^{(t)}$ where $t_u^{(t)} \sim \text{Ber}(1 - \epsilon_u)$, i.e., $\text{Pr}(t_u^{(t)} = 0) = 1 - \epsilon_u$ and $\text{Pr}(t_u^{(t)} = 1) = \epsilon_u$. The moment generating function (MGF) of Ber$(1 - \epsilon_u)$ can be written as

$$M_u(t) = \mathbb{E}[e^{t u}] = \epsilon_u + (1 - \epsilon_u) e^t.$$ (12)

where $x = K_t$ and $K_t$ denotes the first and second differentiation of $K(t)$ relative to $t$, respectively.

Using the following lemma, one can easily calculate the moments and distribution of the CN’s update (11).

**Lemma 1:** Consider random variables $O$ and $H$ where $H$ either has no mass at 0 (discrete) or has support $[0, \infty)$. The approximation of the first and second moments of the distribution of $G = O/H$ using the second-order Taylor expansion can be written as

$$E(G) \approx E(O/H) \approx \frac{\mu_O}{\mu_H} - \frac{\text{cov}(O, H)}{\mu_H^2} + \frac{\mu_O \sigma_H}{\mu_H^3}$$

where $\mu_O$ and $\sigma_O$ are the first and second moments of $O$, respectively, and $\text{var}(O/H) \approx \left( \frac{\mu_O}{\mu_H} \right)^2 \frac{\sigma_O^2}{\mu_H^3} - 2 \frac{\mu_O \sigma_H}{\mu_H^3} + \frac{\sigma_H^2}{\mu_H^4}$.

**Proof:** Please refer to [27] and [28].

For the special case of mIoT, the distribution of (11) can be approximated with a Cauchy distribution.

**Lemma 2:** In the mIoT scenario, the global updates (11) can be approximated with a standard Cauchy distribution.

**Proof:** Let $\{I_1, \ldots, I_U, \ldots\}$ be a sequence of independent, but not necessarily identically distributed random variables, each with an expected value of $1 - \epsilon_u$ and variance of $\epsilon_u(1 - \epsilon_u)$. Taking into account $S_u^2 = \sum_{u=1}^{U} \epsilon_u(1 - \epsilon_u)$, based on Lyapunov’s central limit theorem (CLT), as $U$ goes to infinity, the distribution of $(1/S_u) \sum_{u=1}^{U} (I_u - 1 + \epsilon_u)$ converges to a standard normal distribution $N(0, 1)$. In other words, the numerator and denominator of (11) can be approximated by the zero mean Gaussian distributions. The distribution of the ratio of two independent random variables that both follow a Gaussian distribution with zero mean takes on the shape of a Cauchy distribution. As a result, a Cauchy distribution characterizes the global update in mIoT scenarios when the CN lacks memory.

Now, let us focus on the convergence of FL in the presence of communication errors, when the CN does not have memory.
mass function \((\text{as \cite{20, 29}})\) since \(\epsilon > 0\) \(\forall u \in \mathcal{U}\), and the channels are independent, the probability mass function (pmf) of global parameter can be calculated as \cite{20, 29}:

\[
\Pr \left\{ \mathbf{w}^{(t+1)} = \frac{\sum_{i=1}^{U} I_u^{(t)} u_{i}^{(t)}}{\sum_{i=1}^{U} I_u^{(t)}} \right\} = \prod_{i=1}^{U} \epsilon_{u_i}^{1-I_{u_i}^{(t)}} (1-\epsilon_{u_i})^{I_{u_i}^{(t)}}. \tag{18}
\]

One can easily see that the random characteristic of erasure occurrences is critical to (18), and results in fluctuation of the expected loss. Even though the instantaneous loss of the FL algorithm with erroneous communications without CN memory fluctuates, we will prove that the expected value of the loss converges.

We assume that \(F_u()\), for all \(u \in \{1, \ldots, U\}\), is strongly convex with a parameter \(\mu\) and \(L\)-smooth. For function \(F_u : \mathbb{R}^d \rightarrow \mathbb{R}\) that is \(L\)-smooth \(\forall v_1,v_2 \in \mathbb{R}^d\) \cite{30}

\[
F_u(v_1) \leq F_u(v_2) + \nabla F_u(v_2)^T (v_1 - v_2) + \frac{L}{2} \|v_1 - v_2\|^2. \tag{19}
\]

One can easily show that, \(F(\mathbf{w})\) is also \(L\)-smooth \cite{20}, where

\[
F(\mathbf{w}) = \frac{1}{U} \sum_{u=1}^{U} F_u(\mathbf{w}). \tag{20}
\]

**Lemma 3:** Assuming \(F()\) is convex and \(L\)-smooth \(\forall \mathbf{w} \in \mathbb{R}^d\), and \(F(\mathbf{w}) = (1/U) \sum_{u=1}^{U} (1/D_u) \sum_{i=1}^{D_u} f(\mathbf{w}, x_{u_i}^{(t)}, y_{u_i}^{(t)})\), there exist constants \(M\) and \(N\) such that

\[
\|\nabla f(\mathbf{w}, x, y)\|^2 \leq M\|\nabla f(\mathbf{w})\|^2 + N
\]

for all \((x, y) \in \{(x_{u_i}^{(t)}, y_{u_i}^{(t)}) : u \in \mathcal{U} \text{ and } i \in \mathcal{D}_u\}\).

**Proof:** The proof of existence of such upper bound has been given in \cite[Th. 4.7]{31}.

It is worth mentioning that there are numerous articles that assumed that the stochastic gradients are uniformly bounded which is called bounded dissimilarity, i.e., \(\mathbb{E}[\|\nabla f(\mathbf{w}(t))\|^2] \leq \sigma^2\). However, in \cite{30}, it has been shown that this claim is in contradiction with strong convexity.

**Theorem 1:** Let us consider an FL scenario with \(U\) devices with the data set \(D_u\) of size \(D_u\), where the channel from device \(u\) to the CN is modeled by an erasure channel with erasure probability \(\epsilon_u\). We assume that the loss function \(F(\mathbf{w})\) is strongly convex with a parameter \(\mu\) and \(L\)-smooth. For \(\mu \leq L\), \(M < [D/(\sum_{u=1}^{U} D_u \epsilon_u)]\), and \(\eta^{(t)} = (1/L)\), the expected total loss of the FL algorithm (11) converges as follows:

\[
\mathbb{E}[F(\mathbf{w}^{(t)})] - F(\mathbf{w}^*) \leq CA^t + B \frac{1-A^t}{1-A} \tag{21}
\]

where

\[
A = 1 - \frac{\mu}{L} \left(1 - \frac{M \sum_{u=1}^{U} D_u \epsilon_u}{D}\right) \tag{22}
\]

\[
B = \frac{N}{2L} \sum_{u=1}^{U} D_u \epsilon_u - \frac{D}{D}\tag{23}
\]

\[
C = \mathbb{E}[F(\mathbf{w}^{(0)})] - F(\mathbf{w}^*) \tag{24}
\]

and \(\mathbb{E}(\cdot)\) is the expectation with respect to the packet error.

**Proof:** The proof is provided in Appendix A. \(\blacksquare\)

**Remark 1:** Since \(A < 1\), when \(t\) goes to infinity, the expected loss will converge to

\[
\mathbb{E}[F(\mathbf{w}^{(\infty)})] \leq F(\mathbf{w}^*) + \frac{B}{1-A}. \tag{25}
\]

Remark 1 shows that as \(t\) goes to infinity, the global loss converges to a ball around the optimal value with a radius of \([B/(1-A)]\). Let us assume that \(D_u = D/U\) and \(\epsilon_u = \epsilon \forall u \in \mathcal{U}\). We then have \(A = 1 - (\mu/L)(1 - M\epsilon)\) and \(B = (N/2L)\epsilon\). We then have

\[
\mathbb{E}[F(\mathbf{w}^{(\infty)})] \leq F(\mathbf{w}^*) + \frac{N\epsilon}{2\mu(1-M\epsilon)} \tag{26}
\]

where \(M < 1/\epsilon\).

**Remark 2:** The gap to the optimal global loss increases with \(\epsilon\), since the radius of the ball around the minimum global loss, i.e., \((N/\epsilon)\epsilon(1-M\epsilon))\), is an increasing function of \(\epsilon\). It is also clear that in the noiseless case, i.e., \(\epsilon = 0\), the expected loss will converge to the optimal global loss.

Fig. 2 shows the global loss of the FL scenario with communication errors when the CN does not have memory. As can be seen, the instantaneous MSE will fluctuate significantly with the increase of error probability \(\epsilon\), whereas the average MSE will converge. Moreover, the gap between the average MSE and the optimal global loss increases with \(\epsilon\), which confirms our finding in Remark 2.
B. FL With Erroneous Communication When the CN Has Memory

In the preceding section, we demonstrated that the expectation of loss converges when the CN lacks memory. Although the expected loss converges, the instantaneous loss, and the global parameters fluctuate, which means that the FL algorithm does not necessarily converge. In the following, we explore two instances in which the CN has memory and demonstrate that even in the presence of communication errors, the FL algorithm converges to the optimal global parameters.

1) CN Caches Users’ Local Parameters: Now, we consider the case where the CN has memory to store past local updates of all users. In such a scenario, the distribution of global gradient is the weighted sum of two Bernoulli random variables. In particular, when \( D_a = D/U \) for all users, we have

\[
\omega^{(t+1)} = \frac{1}{U} \sum_{u=1}^{U} \left( \alpha_u I_u^{(t)} + \omega_{u_{-1}} - I_u^{(t)} \right)
\]

where \( I_u^{(t)} \sim \text{Ber}(1 - \epsilon_u) \). Having the MGF of \( I_u^{(t)} \) as \( M_u(t) \), the MGF of \( \alpha I_u^{(t)} + \beta \), which is a linear transformation of \( I_u^{(t)} \), can be written as \( e^{\beta t} M_u(at) \). First, we find \( M_u(t) \) as follows:

\[
M_u(t) = \mathbb{E}\left[ e^{t I_u} \right] = \epsilon_u + (1 - \epsilon_u) e^t.
\] (28)

Next, we calculate the MGF of the weighted sum of the Bernoulli random variables, i.e., \( \omega^{(t+1)} \), as

\[
M(t) = \prod_{u=1}^{U} \mathbb{E}\left[ e^{t (\alpha_u I_u^{(t)} + \omega_{u_{-1}} - I_u^{(t)})} \right]
\]

\[
= \prod_{u=1}^{U} e^{t \omega_{u_{-1}}/2} M_u\left( t \frac{\omega_u - \omega_{u_{-1}}}{U} \right)
\]

\[
= \prod_{u=1}^{U} e^{t (\omega_{u_{-1}}/2)} \left( \epsilon_u + (1 - \epsilon_u) e^{t \omega_{u_{-1}}/2} \right).
\] (29)

Hence, the CF of \( \omega^{(t+1)} \) can be computed as

\[
K(t) = \sum_{u=1}^{U} \left( t \omega_{u_{-1}}/U \right) + \ln \left( \epsilon_u + (1 - \epsilon_u) e^{t \omega_{u_{-1}}/2} \right).
\] (30)

Applying the saddle point approximation, the PDF of \( \omega^{(t+1)} \) can be estimated as

\[
f(t) \approx 2 \pi \sum_{u=1}^{U} \frac{\epsilon_u (1 - \epsilon_u) \Psi_u^{(t+1)^2} e^{\Psi_u^{(t+1)}}}{(\epsilon_u + (1 - \epsilon_u) e^{\Psi_u^{(t+1)}})^{0.5}}
\]

\[
\times \prod_{u=1}^{U} \left( \epsilon_u + (1 - \epsilon_u) e^{\Psi_u^{(t+1)}} \right) \exp \left( -t \left( 1 - \epsilon_u \right) \Psi_u^{(t+1)} e^{\Psi_u^{(t+1)}} \right)
\] (31)

where \( \Psi_u^{(t)} = ((\omega_u^{(t)} - \omega_{u_{-1}})/U) \).

In an FL scenario with erroneous communication, where the CN stores the local parameters of users, the global parameter \( \omega^{(t)} \) converges to the optimal global parameter \( \omega^* \) by increasing the number of communications between the CN and users to be arbitrarily large \([20, Th. 1]\). In an mIoT setup, however, where the density of IoT devices might exceed 20,000 per square kilometer, assuming such memory at the CN is impractical \([32]\). To tackle this scaling problem, we will propose a model in which, given a small number of memory units at the CN, the global parameter \( \omega^{(t)} \) converges to its optimum value, \( \omega^* \).

2) CN Caches Global Parameters: Assume the server has limited memory; hence, it will store the previous \( m \) global updates. In such a scenario, the distribution of global gradient is again the weighted sums of the Bernoulli distributions. For the case of a CN with \( m \) memory units, we can write

\[
\omega^{(t+1)} = \frac{1}{U} \sum_{u=1}^{U} \left( \omega_u^{(t)} I_u^{(t)} + (1 - I_u^{(t)}) \sum_{i=0}^{m-1} \alpha_t - \omega^{(t-i)} \right)
\]

\[
= \sum_{i=0}^{m-1} \alpha_t - \omega^{(t-i)} + \frac{U}{U} \left( \omega_u^{(t)} - \sum_{i=0}^{m-1} \alpha_t - \omega^{(t-i)} \right)
\] (32)

where \( \sum_{i=0}^{m-1} \alpha_t = 1 \), and \( \omega_u^{(t)} \sim \text{Ber}(1 - \epsilon_u) \). By calculating the MGF and CGF of \( Y^{(t)} \), one can see that the saddle point approximation of the distribution is the same as \( (31) \) with \( \Psi_u^{(t)} = (\omega_u^{(t)} - \sum_{i=0}^{m-1} \alpha_t - \omega^{(t-i)}/U) \).

In what follows, we show that the FL algorithm in the presence of communications errors, converges to the optimal parameters, when the CN has memory, even a single unit.

Theorem 2: Consider an FL scenario with \( U \) devices, where the channel between each device and the CN is modeled by an erasure channel with erasure probability \( \epsilon \), and the local loss function \( F_x(a) \) at device \( u \) is convex and \( L \)-smooth. Let us assume \( ||\nabla F(x) - \nabla F(y)||_2 \geq \mu ||x - y||_2 \), for all \( x, y \in \mathbb{R}^d \), where \( F(x) = (1/U) \sum_{u=1}^{U} F_u(x) \). Considering \( \delta_t = ||\omega^{(t)} - \omega^*||_2 \), where \( \omega^* = \arg \min_{\omega} F(\omega) \), and \( \delta_{t+1} = (1/(t + 1)) \sum_{i=0}^{t} \delta_t \), for the FL algorithm \( (10) \) with single memory unit, when \( \epsilon \leq (\mu/2L) \) and \( \eta = (1/L) \), \( \delta_t \) is upper bounded by

\[
\delta_t \leq \frac{F(0) - F(\omega^*)}{t \beta^2}, \quad t > 0
\] (33)

where \( \beta^2 = (\mu^2/2L) - 2L \epsilon^2 \).

Proof: The proof is provided in Appendix B.

Remark 3: According to Theorem 2 as the number of iterations \( t \) increases, the gap between the global parameter \( \omega^{(t)} \) and the global minima \( \omega^* \) decreases (since otherwise if \( \delta_t \geq \epsilon_0 \), where \( \epsilon_0 > 0 \), \( \delta_t \) will be always bounded above \( \epsilon_0 \)).

We emphasize that Theorem 2 and Remark 3 do not state that \( \delta_t \) is a decreasing function of \( t \); instead, it shows \( \omega^{(t)} \) converges to \( \omega^* \) when \( t \) is sufficiently large even in the presence of communication errors.

IV. NUMERICAL RESULTS

We start by investigating the performance of FL with short-packet communications over the Rayleigh fading channel. We choose a message length of \( k = 100 \) bits due to the fact that
the normal approximation for AWGN channels is relatively good for packet lengths of $k \geq 100$ bits and $R \geq 0.5$ [22]. We consider there are $U = 10$ users in the mIoT scenario. As shown in Fig. 3(d), the non-i.i.d. data sets are created using the nonlinear model $y = x^2 + \lambda$, where $\lambda \sim \mathcal{N}(0, 5)$. The data set for each Figs. 3 and 4 has been regenerated.

We consider the symbol duration as the time unit. We assume for all approaches the devices use the same modulation and that the symbol duration is the same. When the rate $R$ is reduced, the packet length increases, and the time required to transmit the packet increases. The duration of each communication round is considered to be equivalent to the packet duration, i.e., $n$ symbols. Thus, for example, 1000 time units correspond to ten rounds for a rate 1 code (assuming one symbol corresponds to 1 bit and each message is 100 bits, so ten messages of 100 bits can be communicated) but only five rounds for a rate 1/2 code (as 200 symbols are sent in each round). Furthermore, the Monte Carlo simulations are used in short-packet communications, i.e., we ran 100 simulations and averaged the results. We assume all users are allocated with orthogonal radio resources; therefore, their simultaneous transmissions do not interfere with each other.3

Figs. 3 and 4 show the global loss of FL based on the coding-rate $R$ in short-packet erroneous communication, when the CN stores local update of IoT devices and global updates, respectively. Considering fading channel, given $R$ and SNR, we use normal approximation bound (3) to calculate the probability of an erroneous communication. Figs. 3(c) and 4(c) illustrate that in the high-SNR regime, increasing the code rate results in a lower convergence time. Because the users’ power is sufficient to overcome the fading and noise, devices with higher rates will communicate more often than those with lower rates over a given learning period, leading the MSE to converge faster. In the low-SNR regime, lower code rates are preferable for a shorter convergence time since we must compensate for the channel noise and fading by adding more parity bits to the message. However, the minimum code rate does not necessarily result in the minimum MSE. Figs. 3(b) and 4(b) demonstrate that for low SNRs, a coding rate of $R = 0.66$ yields the best performance. It is important to note that, in this scenario, the performance of FL when the code rate is $R = 0.2$ is significantly inferior to that of $R = 1$. When the SNR is further reduced, as illustrated in Figs. 3(a) and 4(a), the optimal coding rate $R$ is decreased.4

3A more general case in which radio resource allocation, with the possibility of nonorthogonal transmissions, can be considered to optimize the learning accuracy or the convergence time. This is, however, out of the scope of this work and will be considered by the authors in future works.

4We also investigated the impact of adding new data to the data set of each user at each communication round. Given that IoT measurements do not drastically change over time, the gradient value at each round remains relatively stable. Consequently, this does not noticeably affect the results. Therefore, the results have not been added to this article.
Fig. 5 shows the MSE performance for various coding rates and multiple SNRs, assuming a fixed convergence time. It is evident that, for SNRs greater than 1 dB, the highest coding rate results in the minimum MSE. Nevertheless, as the SNR decreases, the optimal coding rate changes, necessitating careful selection of the coding rate based on the SNR regime.

Fig. 6 shows the loss performance of erroneous short-packet communication as a function of $R$ and SNR. We set the time to $1500 T_s$, where $T_s$ is the symbol duration. As can be seen, when the SNR is low, for a descent loss performance, we need lower $R$ to mitigate the impact of high noise and fading. However, in high-SNR scenarios, it is better to increase the code rate $R$. This is due to the fact that when $R$ is low, the packet length $n$ increases, and subsequently the number of communication rounds between devices and the CN decreases, leaving users with little time to achieve ideal performance. Fig. 6 clearly demonstrates that when $R$ grows, the number of communications for given times will increase and as a result, the MSE will decrease.

When the CN does not have enough storage capacity, instead of storing the local parameters of devices, it will store the past $m$ global updates (Section II-B2b). Fig. 7 shows the impact of the available memory when the CN uses memory to cache the previous $m$ global updates. Here, we considered all the past stored global updates having similar weight ($\alpha = 1/m$). As can be seen, the performance converges with just one unit of memory ($m = 1$), and no fluctuations are visible. Additionally, it is evident that the implementation of multiple memory units ($m > 2$) and utilizing equal weight averaging does not significantly enhance the performance of the system. This finding shows that beyond a certain threshold, the utilization of additional memory units may not yield substantial improvement.
Fig. 8 examines the effect of averaging model on the FL performance when we apply an FL scheme where the CN caches the past $m = 6$ global updates. As is shown, we compare the performance of equal-weighted averaging with the exponentially weighted moving average (EWMA). The EWMA can be written as

$$\omega^{(i)} = \alpha \sum_{i=1}^{m} (1 - \alpha)^{i-1} \omega^{(i-1)}. \quad (34)$$

A larger $\alpha$ accelerates the process of discounting earlier observations. As can be seen, the EWMA with $\alpha \geq 0.5$ has better performance compared to equal-weighted averaging due to the fact that it gives higher weight to the recent global parameter. Our findings show that for $\alpha > 0.95$, the performance increase is minimal.

Despite several studies [33], [34], [35] addressing packet erasure in FL over wireless communication through packet retransmission strategies, this approach has significant drawbacks. Primarily, the usage of packet retransmission strategies can result in a prolongation of the FL convergence time. This delay arises as the CN is required to wait for the receipt of erroneous packets prior to computing the average parameters from all users. Furthermore, there is an inherent lack of assurance that the retransmitted packet will be decoded accurately. This factor of uncertainty becomes markedly significant in a scenario of low SNR, where incorrect retransmissions can incite variability in the FL convergence. As a consequence, the retransmission methods, as explored in prior works, may not yield consistently reliable performance under conditions of low SNR. Fig. 9 showcases that our proposed model outperforms the retransmission strategies cited in previous studies, regardless of high or low coding rate conditions. Moreover, as observed in Fig. 9(b) where the SNR is set at 2 dB, there is an inherent inconsistency in the MSE performance, despite the single retransmission of the erroneous packet under the retransmission strategy. This observation underscores the potential instability associated with this approach.

V. MASSIVE IoT: A PRACTICAL SCENARIOS

In this section, we consider two real-world scenarios to show the tradeoff between the rate, SNR, and probability of error and the impact on the convergence time and accuracy. We discuss both short- and long-packet communication.

A. Short-Packet Communication

We examine the effectiveness of the proposed FL methods in a practical mIoT scenario. Our study makes use of the data set in [21], which is the measurements of the channel transfer function in the frequency domain using a Rhode & Schwartz ZVB14 vector network analyzer (VNA). This measurement system employs omni-directional antennas at both the CN and IoT devices, with both the transmitter and receiver positioned at a height of 1.5 m. The VNA collects 601 frequency samples at 0.167 MHz intervals in a 100 MHz spectrum beginning from 2.4 GHz. As shown in Fig. 10, 196 IoT devices are positioned on a 14 × 14 grid separated by one wavelength ($\lambda = 12.5$ cm). It is important to note that the real-time measurements were performed with the transmitter and receiver in a completely static environment.

We assume the data of each user can be estimated using

$$g(x, y, f) = \sum_{i=1}^{L_1} \kappa_i f^{-i+1} + \sum_{i=1}^{L_2} \upsilon_i x^{i+1} + \sum_{i=1}^{L_3} \eta_i y^{i+1} \quad (35)$$

where $f$ is the frequency, and $x$ and $y$ are the coordinates of the user. In the case of centralized learning, IoT devices send the whole data to the CN for estimating $\kappa$, $\upsilon$, and $\xi$. This means that the CN needs $601 \times 196 \times (L_1 + L_2 + L_3)$ parameters for fitting. In the case of FL, each user applies the GD on its own data and transmit the learning parameters to the CN. Therefore, the maximum number of parameters transmitted by devices is $196 \times (L_1 + L_2 + L_3)$. We consider $L_1 = 3$, $L_2 = 2$, $L_3 = 2$. Therefore, at each iteration of FL, users have seven coefficients to transmit. Owing to the fact that IoT devices are power limited, we assume that they are using IEEE-754 half precision standard for quantization, therefore, each parameter will be converted to a 16-bit digit that has up to eight floating point number precision. As a result, each user have a packet of 112 bits to transmit to the CN.

The majority of IoT devices use a low-power microcontroller unit in which algorithms with minimal computation complexity can be executed [36], [37]. We assume that users have access to the Particle Electron board, a popular board that has seen extensive use in IoT applications. It runs at a
Fig. 9. Convergence of FL when the CN has memory to store local/global parameters of devices versus the case of retransmission in erroneous communication, for $U = 10$, $\eta = 0.05$, $|D_u| = 100$, $\gamma_0 = 2$ dB, $m = 2$, and one iteration of GD at device is applied. (a) $R = 0.2$. (b) $R = 1$. (c) Data set.

speed of 120 MHz thanks to its ARM Cortex M3 processor and has a max gain of 5 dBi at 2.45 GHz. Particle Electron is equipped with a flexible ultra wideband polymer antenna that covers all common LTE frequencies [38]. Therefore, the transmission of users is assumed to be orthogonal, and they undergo Rayleigh fading. We are interested in determining the overall FL system delay given the SNR and code rate. We further disregard the CN’s communication and processing time since it has a high SNR, a large bandwidth, and a high-computing capacity. Therefore, it is reasonable to conclude that the delay results from the processing and transmission times of individual IoT nodes. Each user’s computation time during local training can be determined using

$$t_{\text{comp}}^u = \frac{\rho \mu_u |D_u|}{C_{C_u}}$$

(36)

where $\rho$ denotes the number of CPU cycles required to process a single sample of data which is set to $10^4$ cycle/sample [39], $C_{C_u}$ is the computational capacity of user $u$ which is $1.2 \times 10^8$ for Particle Electron board, $\mu_u$ is the number of iterations of GD for local training of user $u$ in a communication round, and $|D_u|$ is the number of samples of user $u$. Furthermore, the communication time can be computed using

$$t_{\text{comm}}^u = \frac{L_u}{R_u}$$

(37)

where $L_u$ is the size of model parameters of user $u$ and $R_u$ is the transmission rate from user $u$ to the CN. We use the information of the blue users in Fig. 10(a) for training purposes. To ensure that the data set is non-i.i.d. and imbalanced across all the devices, we assume that each user in each row measures the channel transfer function at a fraction of frequencies and that the size of each user’s data set is a random number between 80 and 120. For the test, we use 30% of the data of green users in each row.

The results are presented in Fig. 11 which shows the convergence of MSE in an mIoT scenario for various coding rates across different SNR regimes. We consider $R = 1$ in the case of FL without error. Fig. 11(a) shows that in the low SNRs, the selection of high coding rates results in a longer convergence time. Conversely, Fig. 11(c) demonstrates that in high SNRs, higher coding rates are recommended for faster convergence. However, as is shown in Fig. 11(b), it is apparent that in the mid-SNR regime, the selection of the code rate requires careful consideration, as neither a high code rate nor a low code rate is optimal.

B. Long-Packet Communication

To investigate the performance of the proposed FL schemes in long-packet communication, we consider image classification and apply the FL to train a neural network using a highly non-i.i.d. data set. In order to compute the PER for long-packet communication across a fading channel, given $R$ and the transmitted SNR $\gamma_0$, we use (2) and check to see if the instantaneous received SNR is above the SNR threshold required for error-free communication. We consider the MNIST digits data set, which consists of handwritten images of each number 0 to 7. We considered eight IoT devices ($U = 8$), each of them with 1000 images of one of the numbers between 0 and 7. We employed a MATLAB parallel pool with eight workers and allocated 70% of the data set to training, 15% to test, and 15% to validation. In our simulation, we deployed a CNN model consisting of nine layers, designed to accommodate our MNIST data set. This model includes an image input layer,
the accuracy of FL schemes for error-free and erroneous communication, which correlates to our data set’s classes, and a soft-max layer size of 5. The model concludes with a fully connected layer depth scales from 32 to 64, with each layer having a series of alternating convolutional layers and ReLU activation layers for feature extraction, and max pooling layers for a single memory unit has a significant effect on the performance of FL. While the communication errors are deleterious to the reliability of the packets, the effect can be easily compensated by reusing past local or global parameters, in case of communication errors. This is of significant importance for mIoT systems, as one can relax the reliability requirement, and still achieve the desired level of accuracy within the required time.

**APPENDIX A**

**PROOF OF THEOREM 1**

The global model update can be written as

\[
\omega^{(t+1)} = \omega^{(t)} - \eta^{(t)} \sum_{u=1}^{U} D_u \nabla F_u(\omega^{(t)}) I_u^{(t)} \sum_{u=1}^{U} D_u I_u^{(t)}
\]

that is equivalent to \( \omega^{(t+1)} = \omega^{(t)} - \eta^{(t)} (\nabla F(\omega^{(t)}) - Q) \), where

\[
Q = \nabla F(\omega^{(t)}) - \sum_{u=1}^{U} D_u \nabla F_u(\omega^{(t)}) I_u \sum_{u=1}^{U} D_u I_u.
\]

Assuming that \( v_1 = \omega^{(t+1)} \) and \( v_2 = \omega^{(t)} \), we have \( v_1 - v_2 = -\eta^{(t)} (\nabla F(\omega^{(t)}) - Q) \). Subsequently, since \( F(\cdot) \) is \( L \)-smooth, based on Lipschitz’s gradient (19), we have

\[
F(\omega^{(t+1)}) \leq F(\omega^{(t)}) + \nabla F(\omega^{(t)})^T ( -\eta^{(t)} (\nabla F(\omega^{(t)}) - Q)) + \frac{L}{2} \| -\eta^{(t)} (\nabla F(\omega^{(t)}) - Q) \|_2^2.
\]

Expanding the aforementioned equation, we have

\[
F(\omega^{(t+1)}) \leq F(\omega^{(t)}) - \eta^{(t)} \| \nabla F(\omega^{(t)}) \|_2^2 + \eta^{(t)} \nabla F(\omega^{(t)})^T Q + \frac{L \eta^{(t)}^2}{2} \| \nabla F(\omega^{(t)}) \|_2^2 - L \eta^{(t)}^2 \nabla F(\omega^{(t)})^T Q + \frac{L \eta^{(t)}^2}{2} \| Q \|_2^2.
\]

By fixing the learning rate to \( \eta^{(t)} = (1/L) \), we have

\[
F(\omega^{(t+1)}) \leq F(\omega^{(t)}) - \frac{1}{2L} \| \nabla F(\omega^{(t)}) \|_2^2 + \frac{1}{2L} \| Q \|_2^2. \]

Now, we calculate \( \| Q \|_2^2 \).
\[ \frac{1}{D} \sum_{u=1}^{U} \sum_{i=1}^{I_u} \left\| \nabla f(\omega(0), x_i^u, y_i^u) \right\|_2^2 \]

where in step (i), we assumed that \( S^{(0)} = \{ u | u \in \mathcal{U} \text{ and } I_u^{(0)} = 1 \} \) and \( T^{(0)} = \{ u | u \in \mathcal{U} \text{ and } I_u^{(0)} = 0 \} \). This can be further expanded as follows:

\[ \frac{1}{D} \left( \sum_{u \in S^{(0)}} D_u \left( \sum_{i \in T^{(0)}} D_{i}^{u} \| \nabla f(\omega(0), x_i^u, y_i^u) \|_2^2 \right) + \left( \sum_{u \in T^{(0)}} D_u \left( \sum_{i \in T^{(0)}} D_{i}^{u} \| \nabla f(\omega(0), x_i^u, y_i^u) \|_2^2 \right) \right) \right)^2 \]

which follows from the fact that \( \| I_{u_i}^{(0)} \| = 1 - \epsilon_u \). Since the global loss function is strongly convex with a parameter \( \mu \), according to \cite[Lemma 2.1]{31}, we have

\[ \| \nabla F(\omega(0)) \|_2^2 \geq 2\mu (F(\omega(0)) - F(\omega^*)) \]
It is easy to show that $\|\nabla F(x) - \nabla F(y)\|_2 \geq \mu \|x - y\|_2$ for all $x, y \in \mathbb{R}^d$, we have

$$F(\omega^{(t+1)}) \leq F(\omega^{(t)}) - \frac{\mu^2}{2L} \omega^{(t)} + \frac{L \epsilon^2}{2} \omega^{(t)} - L \epsilon \omega^{(t)} - \omega^* \right) \leq 2(\delta_t + \delta_{t-1}).$$

It is easy to show that $\|\omega^{(t+1)} - \omega^{(t)}\|_2 \leq 2(\delta_t + \delta_{t-1})$. We can further simplify (53) as follows:

$$F(\omega^{(t+1)}) \leq F(\omega^{(t)}) + \left( L \epsilon^2 - \frac{\mu^2}{2L} \right) \delta_t + \frac{L \epsilon^2}{2} \delta_{t-1}$$

Summing up both sides over $t = 1, \ldots, k$, and using telescopic cancellation, we have

$$F(\omega^{(k+1)}) \leq F(\omega^{(0)}) + \left( 2L \epsilon^2 - \frac{\mu^2}{2L} \right) \sum_{t=1}^{k-1} \delta_t$$

where in (55), we assumed that the first global update, i.e., when $t = 1$, $\omega^{(1)}$ is calculated without any communications error. That is, $F(\omega^{(1)}) \leq F(\omega^{(0)}) - (\mu/2L) \delta_0$. Assuming that $\epsilon \leq (\mu/2L)$, we have $(\mu^2/2L) - 2L \epsilon^2 < (\mu^2/2L) - L \epsilon^2$. Therefore, (56) can be simplified as follows:

$$F(\omega^{(k+1)}) \leq F(\omega^{(0)}) \beta^2 (k + 1) \tilde{\delta}_{k+1}$$

where $\beta^2 = (\mu^2/2L) - 2L \epsilon^2$. By rearranging the above inequality, we have

$$\tilde{\delta}_{k+1} \leq \frac{F(\omega^{(0)}) - F(\omega^{(k+1)})}{(k + 1) \beta^2}.$$
[16] C. Yue et al., “NOMA joint decoding based on soft-output ordered-statistics decoder for short block codes,” in Proc. IEEE Int. Conf. Commun., vol. 22, pp. 2163–2168, May 2022.

[17] M. M. Amiri and D. Gündüz, “Federated learning over wireless fading channels,” IEEE Trans. Wireless Commun., vol. 19, no. 5, pp. 3546–3557, May 2020.

[18] H. Zhang and L. Hanzo, “Federated learning assisted multi-UAV networks,” IEEE Trans. Veh. Technol., vol. 69, no. 11, pp. 14104–14109, Nov. 2020.

[19] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, “A joint learning and communications framework for federated learning over wireless networks,” IEEE Trans. Wireless Commun., vol. 20, no. 1, pp. 269–283, Jan. 2021.

[20] M. Shirvaniomohammadi, A. Salari, Y. Gao, and A. Guha, “Federated learning with erroneous communication links,” IEEE Commun. Lett., vol. 26, no. 6, pp. 1293–1297, Jun. 2022.

[21] M. I. AlHajri, N. T. Ali, and R. M. Shubair, “Classification of indoor environments for IoT applications: A machine learning approach,” IEEE Antennas Wireless Propag. Lett., vol. 17, no. 12, pp. 2164–2168, Dec. 2018.

[22] Y. Polynianski, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, May 2010.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[24] B. Yu, Y. Cai, and D. Wu, “Joint access control and resource allocation for short-packet-based mMTC in status update systems,” IEEE J. Sel. Areas Commun., vol. 39, no. 3, pp. 851–865, Mar. 2021.

[25] S. Lv, X. Xu, S. Han, X. Tao, and P. Zhang, “Energy-efficient secure short-packet transmission in NOMA-assisted mMTC networks with relaying,” IEEE Trans. Veh. Technol., vol. 71, no. 2, pp. 1699–1712, Feb. 2022.

[26] R. W. Butler, Saddlepoint Approximations With Applications, vol. 22, Cambridge, U.K.: Cambridge Univ. Press, 2007.

[27] A. Stuwart and K. Ord, Kendall’s Advanced Theory of Statistics, London, U.K.: Edward Arnold, 1994.

[28] R. C. Eldred-Johnson and N. L. Johnson, Survival Models and Data Analysis, vol. 110. New York, NY, USA: Wiley, 1980.

[29] D. N. Ayyala, “High-dimensional statistical inference: Theoretical development to data analytics,” Handbook Stat., vol. 43, pp. 289–335, May 2020.

[30] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtärik, K. Scheinberg, and M. Takáč, “SGD and Hogwild! Convergence without the bounded gradients assumption,” in Proc. Int. Conf. Mach. Learn., vol. 80, 2018, pp. 3750–3758.

[31] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” SIAM Rev., vol. 60, no. 2, pp. 223–311, 2018.

[32] “Ericsson mobility report.” Ericsson. 2017. [Online]. Available: https://www.ericsson.com/en/local-reports/mobility-report/documents/2017/ericsson-mobility-report-2017.pdf

[33] X. Xu, S. Liu, and G. Yu, “Adaptive retransmission design for wireless federeated edge learning,” ZTE Commun., vol. 21, no. 1, pp. 3–14, 2023.

[34] F. Foulakas, “Federated-learning-driven radio access networks,” IEEE Wireless Commun., vol. 29, no. 4, pp. 48–54, Aug. 2022.

[35] H. Hellström, V. Fodor, and C. Fischione, “Federated learning over-the-air by retransmissions,” IEEE Trans. Wireless Commun., early access, Apr. 26, 2023, doi: 10.1109/TWC.2023.3268742.

[36] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Fortivesi, and F. Kawar, “Squeezing deep learning into mobile and embedded devices,” IEEE Pervasive Comput., vol. 16, no. 3, pp. 82–88, Jul. 2017.

[37] V. Prutyanov, N. Melentiev, D. Lopatkin, A. Menshchikov, and A. Somov, “Developing IoT devices empowered by artificial intelligence: Experimental study,” in Proc. IEEE Global IoT Summit (GlOSS), 2019, pp. 1–6.

[38] “Electron datasheet particle.” Mar. 2022. [Online]. Available: https://docs.particle.io/reference/datasheets/e-series/electron-datasheet.

[39] Z. Zhao, J. Xia, L. Fan, X. Lei, G. K. Karagiannidis, and A. Nallanathan, “System optimization of federated learning networks with a constrained latency,” IEEE Trans. Veh. Technol., vol. 71, no. 1, pp. 1095–1100, Jan. 2022.

[40] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. Adv. Intell. Stat., vol. 54, 2017, pp. 1273–1282.

[41] M. P. Friedlander and M. Schmidt, “Hybrid deterministic-stochastic methods for data fitting,” SIAM J. Sci. Comput., vol. 34, no. 3, pp. A1380–A1405, 2012.

Ayoob Salari (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2014, and the M.Sc. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the Centre for IoT and Telecommunications, University of Sydney, Camperdown, NSW, Australia.

His research interests include channel estimation and signal detection, federated learning, and machine learning-based grant-free access for massive IoT communication over 6G networks.

Sarah J. Johnson (Senior Member, IEEE) received the B.E. degree in electrical engineering and the Ph.D. degree from the University of Newcastle, Callaghan, NSW, Australia, in 2000 and 2004, respectively.

She held a Postdoctoral Fellowship with NICTA, Australia’s Information and Communications Technology Research Center of Excellence, an Australian Research Council Postdoctoral Research Fellowship, and an Australian Research Council Future Fellowship. She is currently a Professor of Electrical Engineering and the Deputy Head with the School of Engineering, University of Newcastle. Her research interests include channel coding, machine-type communications, and IoT.

Branka Vucetic (Life Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Belgrade, Belgrade, Serbia.

She is an Australian Laureate Fellow, a Professor of Telecommunications, and the Director of the Centre for IoT and Telecommunications, University of Sydney, Camperdown, NSW, Australia. Her current research work is in wireless networks and Industry 5.0. In the area of wireless networks, she works on communication system design for 6G and wireless AI. In the area of Industry 5.0, her research is focused on the design of cyber–physical human systems and wireless networks for applications in healthcare, energy grids, and advanced manufacturing.

Prof. Vucetic is a Fellow of the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science.

Mahyar Shirvaniomohammadi (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from The University of Tehran, Tehran, Iran, in 2008, the M.Sc. degree (Hons.) in electrical engineering from the Sharif University of Technology, Tehran, in 2010, and the Ph.D. degree in electrical engineering from The University of Sydney, Camperdown, NSW, Australia, in 2015.

He is currently a Senior Lecturer with the Centre for IoT and Telecommunications, The University of Sydney. His research interests include coding and information theory, rateless coding, communication strategies for the Internet of Things, and information-theoretic approaches to machine learning.

Dr. Shirvaniomohammadi received the Best Paper Award for the 2017 IEEE PMRC, The University of Sydney Postgraduate Award and the Norman I Prize, and The 2020 Australian Award for University Teaching. He was selected as one of the Top 50 Young Scientists in the World by the World Economic Forum in 2018 for his contribution to the Fourth Industrial Revolution. He also serves as a Guest Editor for the Entropy, Transactions on Emerging Telecommunications Technologies, and IEEE Journal of Selected Areas in Information Theory. He is a Fellow of the Higher Education Academy.