






Asymptotic Analysis of Federated Learning Under Event-Triggered Communication

Xingkang He , Xinlei Yi , Yanlong Zhao , Karl Henrik Johansson , *Fellow, IEEE*,
and Vijay Gupta , *Fellow, IEEE*

Abstract—Federated learning (FL) is a collaborative machine learning (ML) paradigm based on persistent communication between a central server and multiple edge devices. However, frequent communication of large ML models can incur considerable communication resources, especially costly for wireless network nodes. In this paper, we develop a communication-efficient protocol to reduce the number of communication instances in each round while maintaining convergence rate and asymptotic distribution properties. First, we propose a novel communication-efficient FL algorithm that utilizes an event-triggered communication mechanism, where each edge device updates the model by using stochastic gradient descent with local sampling data and the central server aggregates all local models from the devices by using model averaging. Communication can be reduced since each edge device and the central server transmits its updated model only when the difference between the current model and the last communicated model is larger than a threshold. Thresholds of the devices and server are not necessarily coordinated, and the thresholds and step sizes are not constrained to be of specific forms. Under mild conditions on loss functions, step sizes and thresholds, for the proposed algorithm, we establish asymptotic analysis results in three ways, respectively: convergence in expectation, almost-sure convergence, and asymptotic distribution of the estimation error. In addition, we show that by fine-tuning the parameters, the proposed event-triggered FL algorithm can reach the same convergence rate as state-of-the-art results from existing time-driven FL. We also establish asymptotic efficiency in the sense of Central Limit Theorem of the estimation error. Numerical simulations for linear

regression and image classification problems in the literature are provided to show the effectiveness of the developed results.

Index Terms—Federated learning, asymptotic convergence, event-triggered, stochastic gradient descent, distributed optimization.

I. INTRODUCTION

MACHINE learning (ML) algorithms have now been widely employed in applications as wide as image recognition, natural language processing, automatic speech recognition, and so on. Since data used for training ML models are often generated at edge devices (e.g., cameras, smart phones and wearable devices), centralized machine learning algorithms relying on data collection from all devices at a central server may face some problems: 1) The algorithms are not efficient when the device number is very large, because the computation and storage burden of central servers are heavy; 2) A centralized implementation of the algorithms may not be feasible for privacy and similar concerns. For example, in medical diagnosis, some patients may not like to share their data due to privacy concerns even if such sharing were allowed by regulations; 3) Collecting an enormous amount of raw data (e.g., videos and images) at a central server can substantially increase the communication burden. Distributed learning algorithms have been developed to tackle the above problems.

Federated learning (FL) is a popular distributed learning architecture that enables multiple edge devices holding local data to jointly train a common machine learning model without explicitly exchanging or sharing the data [1], [2], [3], [4], [5]. The delay minimization for FL over wireless communication networks is investigated in [6], [7], [8]. The joint optimization of energy consumption and completion time in federated learning is study in [9]. A typical mathematical model for these works is introduced as follows. Consider a central server and n edge devices, where device j has its local data $\{\xi_j(t)\}_{t=1}^{T_j}$ and a corresponding loss function $f_j(w) \in \mathbb{R}$ is defined as follows

$$f_j(w) = \frac{1}{T_j} \sum_{t=1}^{T_j} f_j(w, \xi_j(t)),$$

where $w \in \mathbb{R}^m$ is the model parameter¹ to train and $f_j(w, \xi_j(t)) \in \mathbb{R}$ yields the loss from inaccurate predictions in using the data $\xi_j(t)$. Common choices of $f_j(w, \xi_j(t))$ are the

¹In this paper, the ‘model parameter’ to train is called ‘model’ for brevity.

Manuscript received 30 August 2022; revised 13 March 2023 and 19 May 2023; accepted 30 June 2023. Date of publication 18 July 2023; date of current version 16 August 2023. This work was supported in part by the ARO of USA under Grant W911NF1910483, in part by the NSF of USA under Grants CBET-1953090 and ECCS-2020246, in part by the AFOSR of USA under Grant FA9550-21-1-0231, in part by the Knut and Alice Wallenberg Foundation, and in part by the Swedish Foundation for Strategic Research. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sangarapillai Lambotharan. (*Corresponding author: Xingkang He.*)

Xingkang He is with Ericsson AB, SE-164 60 Stockholm, Sweden (e-mail: xingkang0715@gmail.com).

Xinlei Yi and Karl Henrik Johansson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: xinlei@kth.se; kallej@kth.se).

Yanlong Zhao is with the Key Lab of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 10010, P. R. China, and also with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China (e-mail: ylzha@amss.ac.cn).

Vijay Gupta is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, IN 47907 USA (e-mail: gupta869@purdue.edu).

Digital Object Identifier 10.1109/TSP.2023.3295734

mean-square error loss in linear regression problems and the cross-entropy loss (also called logarithmic loss) in multi-class classification problems. The server is deployed to aggregate all the models trained at the devices with their individual data and communicate this aggregated model to all devices for the minimization of a global loss function. Specifically, the goal is to design an algorithm through the collaboration between the server and the devices to solve the following optimization problem:

$$w^* = \arg \min_{w \in \mathbb{R}^m} f(w) \triangleq \frac{1}{n} \sum_{j=1}^n f_j(w). \quad (1)$$

The optimum w^* thus corresponds to the globally trained model.

A major advantage of FL is that the data distributions of different devices could be different, i.e., $\xi_j(t)$ may follow different distributions for different j , $j = 1, 2, \dots, n$. For instance, some devices may collect medical diagnosis data while some others may be wearable devices collecting exercising data. A typical FL algorithm involves multiple communication rounds. In each round, all edge devices upload their models trained with individual data to the central server. The central server aggregates the models from the devices and broadcast this aggregated model to the edge devices. Especially for large-scale training models (e.g., deep neural networks), frequently transmitting heavy models between the server and the devices as in traditional FL can impose considerable communication cost. In this paper, we aim to study communication-efficient FL to solve problem (1).

A. Related Work

There are a number of methods in the literature of FL for alleviating communication issues. One category of such methods is to reduce communication bits via data quantization or sparsification [1], [2] in each communication round. Another is to reduce the number of communication instances in each round. For example, in [10], a communication-efficient asynchronous distributed inference method has been proposed for solving convex and nonconvex optimization problems, where a subset of devices are active in each communication round. However, it is required that each edge device should be active once in a few rounds. Under the setting that a subset of devices are randomly chosen to be active, a number of methods have been proposed for FL in [11]. In [12], a distributed stochastic gradient descent (SGD) algorithm has been proposed under the setting that devices upload local gradients to the server in an event-triggered manner and the server broadcasts the fused model to all devices at each iteration. A probabilistic device selection scheme and data quantization method are jointly designed in [13] to achieve communication-efficient FL. Joint data compression and device selection have also been studied in [3], [4], [5]. However, most device selection methods are based on uniform distribution, not taking data heterogeneity into account. In [14], communication-efficient algorithms considering data heterogeneity have been studied.

After receiving an aggregated model from the server, each edge device updates the model based on its own local data

by using a specified optimization method. For problems with enormous training data, instead of gradient-based optimization methods like gradient descent, it is usually more efficient to utilize stochastic optimization methods, such as stochastic gradient descent as well as its extensions and variants (e.g., momentum based SGD, AdaGrad and Adam). For FL, a number of collaborative stochastic optimization methods have been proposed. For example, [15] has studied a distributed stochastic optimization problem over random networks with imperfect communication under convex constraints and established asymptotic convergence properties. We refer to [16] and the references therein for more works in this direction. In addition, there have been some investigations on communication-efficient distributed optimization. A distributed SGD-based algorithm has been designed in [17] for decentralized training of large-scale machine learning models under event-triggered and compressed communication. Based on event-triggered communication, a fully distributed optimization algorithm for second-order continuous-time multi-agent systems has been proposed in [18]. A distributed event-triggered SGD algorithm with decaying step size has been designed in [19] for solving non-convex optimization problems. Moreover, decentralized optimization under event-triggered communication have been studied in [20], [21].

B. Contributions

Despite the above results from the literature, algorithm design and asymptotic convergence analysis on the event-triggered SGD-based FL have not been well studied under the following setting: 1) the step size and triggering thresholds are not parameterized; and 2) the triggering thresholds are not coordinated among all edge devices. Advantages of 1) include: a) It is possible to use non-continuous decay forms. For example, one can use piece-wise-decay step size, which means the step size decays once for every certain steps; b) In certain applications involving bit constraint, such as signal processing units (e.g., CPU and GPU) or communication channel constraint, fixed-point step size can work, but the floating-point step size with a parameterized form may not work; c) For different applications, one can employ specific forms of step size with several parameters and tune the parameters for achieving desirable performance. However, in the literature, it is usually assumed that the step size has concrete forms, such as $\eta(t) = \frac{a}{(t+1)^\delta}$ with $a > 0$ and $\delta \in (\frac{1}{2}, 1]$ in [19] and $\eta(t) = \frac{2}{\mu(t+K_0)}$ with $K_0 > 0$ and $\mu > 0$ in [13]. An advantage of 2) is that it can allow each edge device adjusts the frequency of uploading models to the central server by considering its own data quality and computing power, which can improve the data utility for better training performance. For example, if a device with higher data quality and computing power, a smaller threshold can be employed, such that this device can upload its model to the central server more frequently than other edge devices.

The contributions of this paper are summarized as follows.

- We propose a novel event-triggered SGD-based FL algorithm under the desired setting of 1) and 2) introduced above.

- Under mild conditions, for the proposed algorithm, we establish asymptotic convergence in expectation and provide an estimate of the convergence rate. By tuning the parameters, the convergence rate of the proposed event-triggered algorithm matches the optimal one from the literature on SGD-based FL with persistent communication.
- By utilizing techniques from stochastic approximation, under mild conditions, we prove the almost-sure (i.e., with probability one) asymptotic convergence for the proposed algorithm and provide an estimate of the convergence rate.
- We prove that despite event-triggered communication and random sampling, the estimation error of the proposed algorithm can achieve asymptotic normality and asymptotic efficiency. To our knowledge, this is the first result for communication-efficient FL.

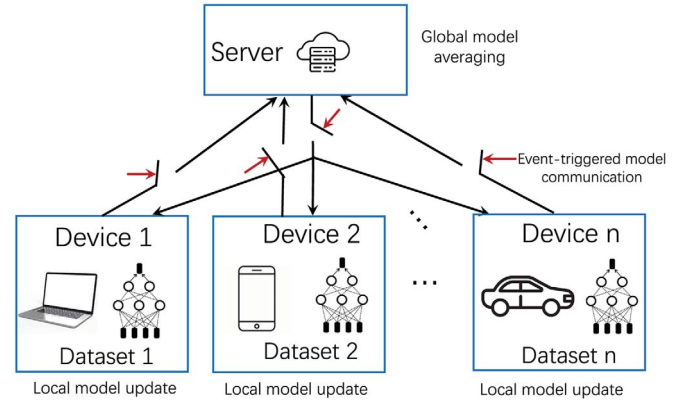


Fig. 1. Event-triggered FL with n devices.

C. Notations and Paper Outline

Notations: \mathbb{N}^+ denotes the set of positive integers and $\mathbb{N} = \mathbb{N}^+ \cup \{0\}$. $\mathbb{E}(x)$ and $\mathbb{E}_\xi(x|\mathcal{F})$ denote the expectation and conditional expectation of random vector x , respectively, where \mathcal{F} is a sigma algebra. $\nabla f(x)$ stands for the gradient of $f(\cdot)$ at x . $\|x\|$ represents the 2-norm of vector x . Given two scalar sequences $\{a_t\}_{t=0}^\infty$ and $\{b_t\}_{t=0}^\infty$, we let $a_t = O(b_t)$ if $\lim_{t \rightarrow \infty} \left| \frac{a_t}{b_t} \right| \leq C < \infty$ and $a_t = o(b_t)$ if $\lim_{t \rightarrow \infty} \frac{a_t}{b_t} = 0$.

Paper outline: The remainder of the paper is organized as follows: Section II is on the main results of the paper, comprising of a novel even-triggered FL algorithm and its asymptotic convergence analysis. In Section III, two numerical simulations are conducted to show the effectiveness of the developed results. The paper is concluded in Section IV. The proofs are given in the Appendix.

II. MAIN RESULTS

In this section, we propose a novel event-triggered FL algorithm and make asymptotic analysis on the convergence of the algorithm in three ways: convergence in expectation, almost-sure convergence, and asymptotic distribution of the estimation error. To our best knowledge, in the literature of event-triggered FL, there have been few results on almost-sure convergence and there has no result on asymptotic distribution.

A. Event-Triggered FL Algorithm

Consider the setting in Fig. 1 with n edge devices and a central server. Each individual edge device can train its model (via neural networks for example) using its own local dataset and the model received from the central server. The central server can aggregate the models of edge devices by model averaging. The communication between the edge devices and the central server follows an event-triggered manner in the sense that each edge device (or the central server) communicates its model only when a certain event is triggered, e.g., the difference between the current model and the last communicated model is larger than a threshold.

Such a setting has several advantages: 1) The edge devices are allowed to be heterogeneous and their local data can follow

different distributions; 2) The data of each edge device is utilized locally without sharing with the central server, which is important to the problems with privacy concerns; 3) The event-triggered communication mechanism exists both in upstream (i.e., from the devices to the server) and downstream (i.e., from the server to the devices), which is more general than the one with only event-triggered communication on the upstream [12].

Let $t \in \mathbb{N}^+$ be the iteration index in the model training. Denote $\nabla f_j(w_j(t), \xi_j(t))$ the stochastic gradient of device j at time t , where $w_j(t)$ represents the estimate of w^* and $\xi_j(t)$ is the stochastically sampled data (via a mini-batch method for example), $j = 1, 2, \dots, n$. Then we propose a new event-triggered FL algorithm as shown in Algorithm 1. The main idea of Algorithm 1 is as follows: At time $t \in \mathbb{N}^+$, the server aggregates the latest received models from all edge devices, and then broadcasts the aggregated model to edge devices if the difference between the model and the last communicated one is larger than a threshold $\mu_a(t) \geq 0$; Each device $j \in \{1, 2, \dots, n\}$ runs a local SGD iteration with its own data for model update, and then uploads the model if the difference between the model and the last-sent-out one is larger than a threshold $\mu_j(t) \geq 0$.

Remark II.1: The differences between Algorithm 1 and existing algorithms include: 1) Different from [12], [19], the triggering thresholds are not coordinated among edge devices, so that each edge device can design its own triggering threshold considering its data quality and computing power, which can improve the data utility for better training performance. For example, if a device with higher data quality and computing power, a smaller threshold can be employed, so that this device can upload its model to the central server more frequently than other edge devices; 2) Unlike [4], [17], [19], the step size $\eta(t)$ and triggering thresholds $\mu_j(t)$, $j \in \{1, 2, \dots, n, a\}$ are not parameterized, so that one can achieve most general fine-tuning on hyper-parameters by considering different parameterizations in different problems.

Remark II.2: The results in the paper can be extended to the setting where the step sizes $\{\eta_i(t)\}_{i=1}^n$ are not coordinated, i.e., $\eta_i(t) \neq \eta_j(t)$ for $i \neq j$, as long as there is a sequence $\eta(t)$ subject to $\lim_{t \rightarrow \infty} \frac{\eta_i(t)}{\eta(t)} = 1$ for any $i = 1, 2, \dots, n$ and the assumptions made in the paper.

In this paper, we make the following assumptions.

Algorithm 1 Event-Triggered Federated Learning (ETFL)

1: **Inputs:** $w_a(0)$ is the initial aggregated model, $\eta(t)$ the step size of edge devices at time t , $\mu_j(t) \geq 0$ and $\mu_a(t) \geq 0$ are the event-triggered thresholds of device j and the server, respectively. Let $\tau_0^a = \tau_0^j = k_a(0) = k_j(0) = 0$.

2: **Output:** Aggregated models $\{w_a(t)\}_{t \geq 1}$ or local models $\{w_j(t)\}_{t \geq 1}$, $j = 1, 2, \dots, n$.

3: **for** $t = 1, \dots$ **do**

4: **for** each edge device j in parallel **do**

5: **if** $t \geq 2$ **then**

6: **if** it receives an updated model $w_a(t-1)$ from the server **then**

7: let $k_a(t-1) = k_a(t-2) + 1$ and $\tau_{k_a(t-1)}^a = t-1$
 // $\tau_{k_a(t)}^a$: last triggering time instant of server till t

8: **else**

9: $k_a(t-1) = k_a(t-2)$ // $k_a(t)$: accumulated triggering times of server till t

10: **end if**

11: **end if**

12: $w_j(t) = w_a(\tau_{k_a(t-1)}^a) - \eta(t-1) \nabla f_j(w_a(\tau_{k_a(t-1)}^a), \xi_j(t-1))$ // Local stochastic gradient update;

13: **if** $\|w_j(t) - w_j(\tau_{k_j(t-1)}^j)\| > \mu_j(t)$ **or** $t = 1$ **then**

14: device j sends its model $w_j(t)$ to the server, and let $k_j(t) = k_j(t-1) + 1$ and $\tau_{k_j(t)}^j = t$

15: **else**

16: $k_j(t) = k_j(t-1)$

17: **end if**

18: **end for**

19: **for** server **do**

20: **if** it receives updated model $w_j(t)$ from edge device j **then**

21: let $k_j(t) = k_j(t-1) + 1$ and $\tau_{k_j(t)}^j = t$ // $\tau_{k_j(t)}^j$: last triggering time instant of device j till t

22: **else**

23: $k_j(t) = k_j(t-1)$ // $k_j(t)$: accumulated triggering times of server till t

24: **end if**

25: $w_a(t) = \frac{1}{n} \sum_{j=1}^n w_j(\tau_{k_j(t)}^j)$ // model aggregation

26: **if** $\|w_a(t) - w_a(\tau_{k_a(t-1)}^a)\| > \mu_a(t)$ **then**

27: the server broadcasts its model $w_a(t)$ to the devices, and let $k_a(t) = k_a(t-1) + 1$ and $\tau_{k_a(t)}^a = t$

28: **else**

29: $k_a(t) = k_a(t-1)$

30: **end if**

31: **end for**

32: **end for**

Assumption II.1: The loss function in (1) is radially unbounded, i.e., $f(w) \xrightarrow{w \rightarrow \infty} \infty$, and has a unique solution w^* , such that $f(w^*) > -\infty$. The gradients of objective functions $\nabla f_i(\cdot) : \mathbb{R}^m \mapsto \mathbb{R}^m$ exist over the whole space and are Lipschitz continuous with Lipschitz constant $L_i > 0$, i.e., it holds that $\|\nabla f_i(w_a) - \nabla f_i(w_b)\| \leq L_i \|w_a - w_b\|$ for any $w_a, w_b \in \mathbb{R}^m$, $i = 1, 2, \dots, n$.

To introduce the next assumption, we let $\mathcal{F}(t) = \sigma(\xi_j(t_s))$, $1 \leq j \leq n$, $0 \leq t_s \leq t$, where $\sigma(\cdot)$ denotes the operator of minimal sigma algebra. In addition, we denote

$$\mathbb{E}_{\xi_j(t)} \{\nabla f_j(w, \xi_j(t))\} = \mathbb{E}_{\xi_j(t)} \{\nabla f_j(w, \xi_j(t)) | \mathcal{F}(t-1)\},$$

which means the expected value taken with respect to the distribution of random variable $\xi_j(t)$ given filtration $\mathcal{F}(t-1)$.

Assumption II.2: Given any $w \in \mathbb{R}^m$ and $t \in \mathbb{N}$, the stochastic gradients $\nabla f_j(w, \xi_j(t))$

1) are conditionally unbiased almost surely, i.e.,

$$\mathbb{E}_{\xi_j(t)} \{\nabla f_j(w, \xi_j(t))\} = \nabla f_j(w), \quad a.s..$$

2) has conditionally bounded second moment, i.e., there is a finite scalar $q_1 > 0$, such that

$$\mathbb{E}_{\xi_j(t)} \left\{ \|\nabla f_j(w, \xi_j(t)) - \nabla f_j(w)\|^2 \right\} \leq q_1, \quad a.s..$$

Remark II.3: The Lipschitz condition on gradients in Assumption II.1 is common in the literature. Comparing with [19], we remove the requirement of the Lipschitz condition on each loss function $f_j(w)$. Assumption II.1 is mild and satisfied with $f(w) = \|w - w^*\|^2$ for instance. Assumption II.2 is common in existing works [4], [21], which allows the data of different devices to satisfy different distributions, reflecting the feature of FL working for non-i.i.d. data.

In the following sections, we will make asymptotic convergence analysis on Algorithm 1 with respect to iterates $w_a(t)$.

B. Convergence in Expectation

In this section, we establish the convergence of Algorithm 1 in the sense of expectation. Recall from Algorithm 1 that $\mu_j(t) \geq 0$ and $\mu_a(t) \geq 0$ stand for the triggering thresholds of device j and the server, respectively. Then denote $\mu_0(t) = \frac{1}{n} \sum_{j=1}^n \mu_j(t) + \mu_a(t)$. We make the following assumption on the step size $\eta(t)$ and triggering parameter $\mu_0(t)$.

Assumption II.3: The following conditions hold:

1. $\eta(t) > 0$, $\sum_{t=0}^{\infty} \eta(t) = \infty$, $\sum_{t=0}^{\infty} \eta^2(t) < \infty$,
 $\lim_{t \rightarrow \infty} \left(\frac{1}{\eta(t+1)} - \frac{1}{\eta(t)} \right) = \eta_0 \geq 0, \forall t \in \mathbb{N}$,
2. $\mu_0(t) > 0$, $\sum_{t=0}^{\infty} \mu_0(t) < \infty$, $\forall t \in \mathbb{N}$,
3. the global loss function $f(w)$ is s -strongly convex, i.e., there is a positive scalar $s > 0$, such that $f(y) \geq f(x) + \nabla f^T(x)(y-x) + \frac{s}{2} \|y-x\|^2$ for any $x, y \in \mathbb{R}^m$.

Remark II.4: Comparing with [13], [19], Assumption II.3 does not confine the specific forms of step size and triggering thresholds. It can be fulfilled with $\eta(t) = \frac{a_1}{a_2 t^l + a_3}$ and $\mu_j(t) = \frac{b_{1,j}}{b_{2,j} t^{l'} + b_{3,j}}$, where $a_i, b_{i,j}, i = 1, 2, 3$, are positive scalars and $l > \frac{1}{2}$ and $r_j > 1$, $j \in \{1, 2, \dots, n, a\}$.

Proposition II.1: Under Assumptions II.1, II.2, and II.3 1)–2), it holds that

$$\sum_{t=0}^{\infty} \eta(t) \mathbb{E} \{ \|\nabla f(w_a(t))\|^2 \} < \infty.$$

Proof: Proof sketch: First, we introduce $\bar{w}(t) = \frac{1}{n} \sum_{j=1}^n w_j(t)$ the average model of all n devices, and then investigate the upper bound of $\mathbb{E}_{\xi(t)} \{f(\bar{w}(t))\} - f(w_a(\tau_{k_a(t-1)}^a))$ by analyzing event-triggered errors. Then we establish an upper bound of $\mathbb{E} \{f(\bar{w}(t))\} - \mathbb{E} \{f(\bar{w}(t-1))\}$ under Assumptions II.1, II.2, and II.3 1)–2). Finally, the conclusion is attained with the help of Lemma A.2.

See Appendix A for the full proof. \square

From Assumption II.3 1)–2) and Proposition II.1, there are $\sum_{t=0}^{\infty} \eta(t) = \infty$ and $\sum_{t=0}^{\infty} \eta(t) \mathbb{E} \{ \|\nabla f(w_a(t))\|^2 \} < \infty$. This result reflects the sub-sequence convergence of $\mathbb{E} \{ \|\nabla f(w_a(t))\|^2 \}$ for any non-convex loss functions $f(\cdot)$. A

stronger convergence result is attained if Assumption II.3 3) is also satisfied.

In case of $\eta_0 = 0$, we let $\frac{1}{0} = \infty$ in the following theorem for notational convenience.

Theorem II.1: Under Assumptions II.1–II.3, let $\mu_0(t) = O(\eta^\rho(t))$ with $\rho > 1$, then for any $\delta \in [0, \min\{\rho_0, s/(2\eta_0)\})$ with $\rho_0 = \min\{1, \rho - 1\}$, it holds that

$$\begin{aligned}\mathbb{E}\{\|w_a(t) - w^*\|^2\} &= o(\eta^\delta(t)), \\ \mathbb{E}\{f(w_a(t))\} - f(w^*) &= o(\eta^\delta(t)),\end{aligned}$$

where $s > 0$ and $\eta_0 \geq 0$ are introduced in Assumption II.3.

Proof: Proof sketch: First, we obtain an estimate on the convergence rate of $\mathbb{E}\{f(\bar{w}(t))\} - f(w^*)$ from the proof of Proposition II.1 and Assumption II.3 -3). Then we prove that the error $\mathbb{E}\{f(w_a(t))\} - f(w^*)$ decays to zero as fast as $\mathbb{E}\{f(\bar{w}(t))\} - f(w^*)$, which leads to the second conclusion. The first conclusion directly follows from the s -strong convexity in Assumption II.3 and the second conclusion.

See Appendix B for the full proof. \square

Remark II.5: Due to $\eta(t) \rightarrow 0$, Theorem II.1 shows the asymptotic convergence with an estimated convergence rate $o(\eta^\delta(t))$, where δ depends on the decaying speed of step size and triggering thresholds. Thus, one can design the step size and triggering thresholds to achieve the desired convergence rate. For example, if the desired minimal convergence rate is $o(\frac{1}{t^s})$, under Assumptions II.1–II.3, the rate can be achieved by letting $\eta(t) = \frac{1}{t}$ and $\mu_0(t) = \frac{1}{t^\rho}$, as long as $\min\{1, \rho - 1, s/2\} > \delta \geq 0$.

Remark II.6: By tuning the hyper parameters $\eta(t)$ and $\{\mu_j(t)\}_{j=1}^n$, the convergence rate in Theorem II.1 can match the optimal SGD convergence rate $O(\frac{1}{t})$ from the literature [13] on FL under persistent communications. For example, let $\eta(t) = \frac{\bar{a}}{t}$ with $\bar{a} > \frac{2}{s}$, $\mu_0(t) = O(\eta^2(t))$, then the convergence rate in Theorem II.1 is $o(\frac{1}{t^\delta})$ for any $\delta \in [0, 1)$. Since δ can be arbitrarily close to 1, the convergence rate is essentially the same as $O(\frac{1}{t})$. Alternatively, we can directly prove that under the above parameter setting, the convergence rate of the Algorithm 1 is $O(\frac{1}{t})$ by referring to the proofs of Theorem II.1 and Lemma A.3. In this paper, we use $o(\cdot)$ to state the main results under a general setting of step size and triggering thresholds. In Theorem II.3, we will show that the time-averaged estimate can achieve the rate $O(\frac{1}{nt})$, i.e., $\mathbb{E}\{f(\hat{w}_a(t))\} - f(w^*) = O(\frac{1}{nt})$.

C. Almost-Sure Convergence

In this section, we establish the almost-sure convergence of Algorithm 1. To proceed, we make the following assumption.

Assumption II.4: The following conditions hold:

1. $\eta(t) > 0$, $\sum_{t=0}^{\infty} \eta(t) = \infty$, $\sum_{t=0}^{\infty} \eta^{2(1-\delta)}(t) < \infty$,
 $\lim_{t \rightarrow \infty} \left(\frac{1}{\eta(t+1)} - \frac{1}{\eta(t)} \right) = \eta_0 \geq 0$, $\forall t \in \mathbb{N}$,
2. $\mu_0(t) > 0$, $\mu_0(t+1) \leq \mu_0(t)$, $\mu_0(t) = O(\eta^{1+\delta}(t))$,
 $\forall t \in \mathbb{N}$,
3. the global loss function is s -strongly convex and twice continuously differentiable,
 where $\delta \in (0, \min\{\frac{1}{2}, \frac{s}{\eta_0}\})$.

Remark II.7: Assumption II.4 is mild. For example, it is fulfilled if $\eta(t) = \frac{1}{t^l}$ and $\mu_0(t) = \frac{1}{t^{l(1+\delta)}}$, $l \in (\frac{1}{2(1-\delta)}, 1]$, and $f(w) = \frac{1}{2}(w - w^*)^\top(w - w^*)$ with $\delta \in (0, \min\{\frac{1}{2}, \frac{1}{\eta_0}\})$. The assumption on the threshold provides a quantitative design principle that the threshold should not decay more slowly than the provided bound. Otherwise, the events in the local devices may not be triggered in time. If so, despite communication saving, the optimization algorithm would be divergent due to too large coordination errors among local devices.

Theorem II.2: The following conclusions hold:

1. Under Assumptions II.1, II.2 and II.3 1)–2), let $\mu_0(t+1) \leq \mu_0(t)$, then it holds that

$$\begin{aligned}w_a(t) &\xrightarrow{t \rightarrow \infty} w^*, \text{ a.s.} \\ f(w_a(t)) &\xrightarrow{t \rightarrow \infty} f(w^*), \text{ a.s.}\end{aligned}$$

2. Under Assumptions II.1, II.2 and II.4, it holds that

$$\begin{aligned}\|w_a(t) - w^*\| &= o(\eta^\delta(t)), \text{ a.s.}, \\ f(w_a(t)) - f(w^*) &= o(\eta^{2\delta}(t)), \text{ a.s.},\end{aligned}$$

where δ is introduced in Assumption II.4.

Proof: Proof sketch: We focus on the convergence of iterates $w_a(t)$, which lead to the almost-sure convergence on $f(w_a(t))$ according to Lemma A.1. First, we obtain the iteration of $w_a(t)$ from the iteration of $\bar{w}(t)$. Then we use Lemma A.5 on stochastic approximation by verifying the conditions on noise, loss function, and step size.

See Appendix C for the full proof. \square

Theorem II.2 establishes almost-sure asymptotic convergence and convergence rate of iterates $w_a(t)$, which has not been well investigated in the related works. According to Assumption II.4 and Theorem II.2, one can tune δ , $\eta(t)$ and $\mu_0(t)$ to reduce the communication while maintaining a desired convergence rate. For example, if the desired minimal convergence rate is $\|w_a(t) - w^*\| = o(\frac{1}{t^s})$, under Assumptions II.1, II.2 and II.4, the rate can be achieved by letting $\eta(t) = \frac{1}{t}$ and $\mu_0(t) = \frac{1}{t^{(1+\delta)s}}$ as long as $\delta \in (0, \min\{\frac{1}{2}, s\})$.

D. Asymptotic Distribution

In this section, we establish the asymptotic distribution of the estimation error $w_a(t) - w^*$, which has not been investigated in the study of communication-efficient FL to the best of our knowledge. To proceed, we introduce some extra conditions as follows.

Assumption II.5: The following two conditions hold:

1. For any $j \in \{1, 2, \dots, n\}$, the stochastic gradient error $\bar{\epsilon}_j(t) = \nabla f_j(w_a(\tau_{k_a(t)}^a), \xi_j(t)) - \nabla f_j(w_a(\tau_{k_a(t)}^a))$ satisfies the following conditions

$$\begin{aligned}\lim_{t \rightarrow \infty} \mathbb{E}\{\bar{\epsilon}_j(t) \bar{\epsilon}_j^\top(t) | \mathcal{F}(t-1)\} &= S_j, \\ \mathbb{E}\{\bar{\epsilon}_i(t) \bar{\epsilon}_j^\top(t) | \mathcal{F}(t-1)\} &= 0, \quad i \neq j, \\ \mathbb{E}\{\|\bar{\epsilon}_i(t)\|^p | \mathcal{F}(t-1)\} &< \bar{q}_1, \quad (2)\end{aligned}$$

where S_j is a finite positive semi-definite matrix, $p > 2$ and $\bar{q}_1 > 0$.

2. The global loss function $f(w)$ is s -strongly convex, $s > 0$, and is three times continuously differentiable and the three-order derivative of $f(w)$ is locally bounded around w^* .

Theorem II.3: Let $\eta(t) = \frac{1}{t^v}$ with $v \in (\frac{2}{3}, 1)$ and $\mu_0(t) = o(\eta(t)^{\frac{3}{2}})$. Under Assumptions II.1–II.2 and II.5, the following conclusions hold:

1. $\frac{1}{\sqrt{\eta(t)}}(w_a(t) - w^*)$ is asymptotically normal, i.e.,

$$\frac{1}{\sqrt{\eta(t)}}(w_a(t) - w^*) \xrightarrow[t \rightarrow \infty]{d} N(0, S),$$

where $S = \int_0^\infty e^{Ht} S_0 e^{H^T t} dt$;

2. $\hat{w}_a(t) := \frac{1}{t} \sum_{i=1}^t w_a(i)$ is asymptotically efficient:

$$\sqrt{t}(\hat{w}_a(t) - w^*) \xrightarrow[t \rightarrow \infty]{d} N(0, \bar{S}),$$

where $\bar{S} = H^{-1} S_0 (H^{-1})^T$. In addition, it holds that

$$\mathbb{E}\{f(\hat{w}_a(t))\} - f(w^*) = O\left(\frac{1}{nt}\right), \quad (3)$$

where $S_0 = \frac{1}{n^2} \sum_{j=1}^n S_j$ and $H = \nabla^2 f(w^*)$.

Proof: Proof sketch: Similar to the proof of Theorem II.2, we obtain the conclusions by applying Lemma A.5 and verifying the required conditions on noise, loss function, and step size.

See Appendix D for the full proof. \square

Remark II.8: As we know that the Central Limit Theorem (CLT) from statistics is important, because it allows us to safely assume that the sampling distribution of the mean is normal in most cases. Theorem II.3 is a generalized CLT, showing that the proposed algorithm can achieve asymptotic normality despite event-triggered communication and noisy data. To the best of our knowledge, this is the first result in the study of communication-efficient FL.

Remark II.9: The convergence result in (3) is the optimal convergence rate from the literature [17]. Comparing with Theorem II.1, this rate reflects the connection with n the number of devices.

Remark II.10: As seen from Theorems II.1–II.3, the threshold $\mu_i(t)$ with a faster decaying speed can lead to a higher convergence rate of Algorithm 1. Thus, if one aims to obtain a higher convergence rate with sufficient budget in communication resource, the threshold $\mu_i(t)$ can be designed to decay very fast or just be zero. In addition, for the scenarios where a subset of devices have higher data quality and computing power, their thresholds can decay faster than others, so that they can upload their models to the central server more frequently than other edge devices. For different applications, one can employ specific forms of the threshold with several parameters as in Remarks II.4 and II.5, and then tune the parameters for achieving desirable performance.

III. NUMERICAL SIMULATIONS

In this section, we conduct two numerical simulations to show the effectiveness of the developed results. The first simulation is on a linear regression problem and the second one is to solve a digit recognition problem with an open dataset.

A. Linear Regression

In this simulation, we consider a linear regression problem with ten devices (e.g., sensors) and a server. Suppose the data of device $j \in \{1, 2, \dots, 10\}$ at time $t \in \{1, 2, \dots, T\}$ is generated as follows: $y_j(t) = H_j w^* + v_j(t)$, where $w^* = [10, -2]^T$ is the unknown parameter or signal to be estimated, $H_j = [1, 2]$ and $v_j(t) \sim N(0, 1)$ for $j = 2, 4, 6, 8, 10$, and $H_j = [-2, 1]$ and $v_j(t) \sim U(-1, 1)$ for $j = 1, 3, 5, 7, 9$. In addition, the noise samples $\{v_j(t)\}$ are drawn independently both in time and in space. The collective loss function to be minimized is $f(w) = \sum_{j=1}^{10} f_j(w)$, where $f_j(w) = \frac{1}{T} \sum_{t=1}^T (y_j(t) - H_j w)^2$. The stochastic gradient of $f_j(w)$ at w is $\nabla f_j(w, \xi_j(t)) = -2H_j^T (y_j(t) - H_j w)$. To employ Algorithm 1, we let the step size $\eta(t) = \frac{1}{10t}$. In order to illustrate the influence of the triggering thresholds to the estimation performance and communication frequency, we consider three different settings on the triggering thresholds: 1) $\mu_a = \mu_j(t) = 0$ for $j \in \{1, 2, \dots, 10\}$; 2) $\mu_a = \mu_j(t) = \frac{3}{10t^{1.3}}$ for $j \in \{1, 3, 5, 7, 9\}$ and $\mu_j(t) = \frac{3}{5t^{1.2}}$ for $j \in \{2, 4, 6, 8, 10\}$; 3) $\mu_a = \mu_j(t) = \frac{3}{10t^{1.1}}$ for $j \in \{1, 3, 5, 7, 9\}$ and $\mu_j(t) = \frac{3}{5t}$ for $j \in \{2, 4, 6, 8, 10\}$. Under the first setting, the devices and the server communicate all the time since the triggering conditions are always satisfied. While, under the other two settings, the devices and the server are expected to communicate intermittently. Note that the triggering thresholds in the third setting decay more slowly than the one in the second setting. We conduct Monte Carlo experiments with M runs by using the following metrics on estimation performance and communication frequency over the network:

$$\text{MSE}(t) = \frac{1}{M} \sum_{s=1}^M \|w_a^s(t) - w^*\|^2,$$

$$\text{communication rate}(t) = \sum_{s=1}^M \frac{nk_a^s(t) + \sum_{j=1}^n k_j^s(t)}{2nMt} \in [0, 1],$$

where n is the device number, $w_a^s(t)$ denotes the parameter estimate of the server at time t in the s -th run, and $k_j^s(t)$ and $k_a^s(t)$ represent the cumulative triggering times of device j and the server till time t in the s -th run, respectively. Let $M = 100$, $t = 1, 2, \dots, 200$, and the initial estimates be $[0, 0]^T$, then we run the simulation with Algorithm 1. The simulation results on MSE and communication rate are respectively given in Fig. 2(a) and 2(b). Fig. 2(a) shows that all the three settings can lead to the convergence of MSE, and the performance under the second setting with fast decaying thresholds is very close to that under the first one (which can be treated the optimal one due to persistent communication). In addition, a derived convergence rate from Theorem II.1, that is $1/(1 + \eta^{0.2}(t))$, is also added for reference. The figure shows that the derived convergence rate is close to the true convergence rate of the proposed algorithm. Fig. 2(b) shows that the event-triggered setting can lead to substantial reduction of communication in the whole process, while maintaining high estimation precision as in Fig. 2(a). In addition, reducing the decaying speed of the thresholds can further reduce the communication. To illustrate the asymptotic distribution of the estimation error $w_a^s(t) - w^*$, we let

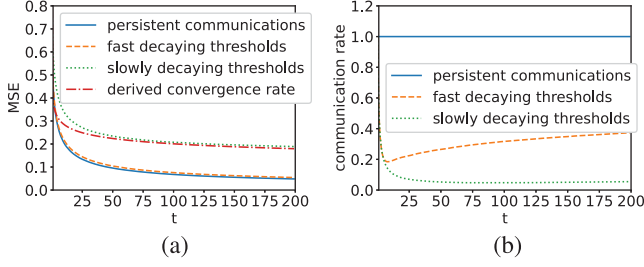


Fig. 2. Performance of the proposed algorithm under three different settings of triggering thresholds. (a) MSEs versus time index. (b) Communication rates versus time index.

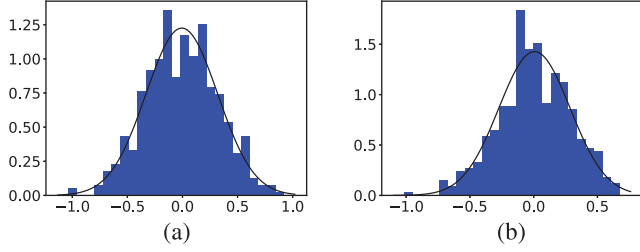


Fig. 3. Distribution of $\frac{w_a^s(t) - w^*}{\sqrt{\eta(t)}}$ at time $t = 1000$. (a) Distribution of the first element. (b) Distribution of the second element.

$M = 500$, $\eta(t) = \frac{1}{10^{3t^{0.5}}}$, $t = 1, 2, \dots, 1000$ and the thresholds be the same as the second setting above. The simulation results on the distributions of the first and second elements of $\frac{w_a^s(t) - w^*}{\sqrt{\eta(t)}}$ at time $t = 1000$ are given in Fig. 3(a) and 3(b) which shows the asymptotic normality conforming with Theorem II.3.

B. Softmax Regression Over MNIST Dataset

In this experiment, we consider an image classification problem based on digit recognition dataset MNIST,² which consists of 60K training samples and 10K testing samples of digits ranging from 0 to 9. To handle this 10-class classification problem, we consider softmax regression with cross-entropy loss. We employ ten devices for model training in the FL, where device i can only have access to samples of digit i for $i \in \{0, 1, \dots, 9\}$, to show the effectiveness of the proposed algorithm in dealing with non-i.i.d. data. Consider four algorithms: Algorithm 1 with step size $\eta(t) = \frac{1}{10^{3t^{0.5}}}$, triggering threshold $\mu_j(t) = \frac{3}{10^{2t^{0.6}}}$, for any $j \in \{1, 2, \dots, 10, a\}$, denoted by ETFL; Algorithm 1 with step size $\eta(t) = \frac{1}{10^{3t^{0.5}}}$, triggering threshold $\mu_j(t) \equiv 0$ for any $j \in \{1, 2, \dots, 10, a\}$, denoted by TTFL, which means the communication between the devices and the server are persistent; The distributed SGD algorithm in [19], denoted by DSGD; The SPARQ-SGD algorithm in [17] but without considering the communication compression step. In addition, the above four algorithms share the following setting: Monte Carlo experiments are conducted with $M = 10$ runs and 200 iterations for each run; A mini-batch method with batch size 600 is used to obtain a stochastic gradient; The initial parameter for model training is $0^{28 \times 28}$.

After running the experiments under the above setting, we provide the results in Fig. 4. Fig. 4(a) and 4(b) respectively

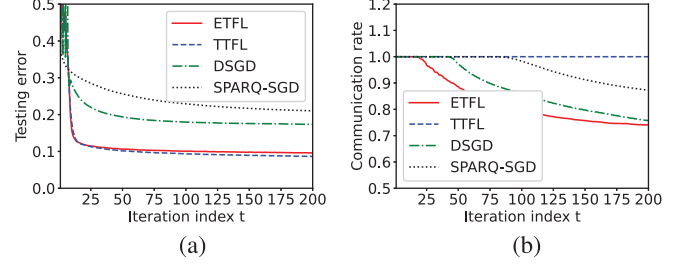


Fig. 4. Performance comparison on the softmax regression problem. (a) Testing errors. (b) Communication rates.

shows testing error and communication rate (defined in the last case) versus iteration index t for the considered three algorithms. It is seen that our proposed ETFL is very close to TTFL in learning performance, while saving more than 20% communication. In addition, comparing with DSGD and SPARQ-SGD, ETFL can lead to better learning performance with less communication.

C. Neural Network Training Over CIFAR-10 Dataset

In this experiment, we consider an image classification problem based on CIFAR-10,³ which is a collection of images that are commonly used to train machine learning and computer vision algorithms. The CIFAR-10 dataset contains 60,000 32×32 color images in 10 different classes.

To handle this 10-class classification problem, we consider a relatively small size neural network via some tools in pytorch. The network is designed as follows: There are two convolution layers whose shapes are 3×32 and 32×64 , respectively, with common kernel size 3, stride 1, and padding 1. After each convolution layer, there is an Relu layer and a 2×2 max-pooling layer. The last two layers are two fully connected layers. The model training is based on the proposed algorithm in this paper. Suppose there are two devices, where the first device has the training data on the first five classes and the other device has data on the other five classes. There are 50,000 training images in total. In addition, there is a server which can communicate the aggregated model with the two devices and evaluate the prediction performance of the model with 10,000 test images during the whole training process. The batch size is set to be 256 in both training and test. The epoch number for training is 50. Furthermore, the cross entropy loss is considered as the criterion. The step size is 0.5 at the first epoch and will be multiplied by 0.95 every epoch. The triggering threshold for the devices and the server is $100/(\text{epoch} + 8)^{1.3}$, epoch = 1, 2, \dots , 50.

After running the experiment under the above setting, we obtain the results in Fig. 5. In Fig. 5(a), the (local) training accuracy for the two devices are depicted as well as the prediction accuracy in the central server. It shows that as epoch increases, the training and prediction accuracies are becoming better and better. It is due to the non-i.i.d. data distribution in the devices that prediction accuracy is not as good as training accuracy. In 5(b), the communication rate, defined as in Subsection III-A, is depicted. From the figure, we see that the communication

²<http://yann.lecun.com/exdb/mnist/>

³<https://www.cs.toronto.edu/~kriz/cifar.html>

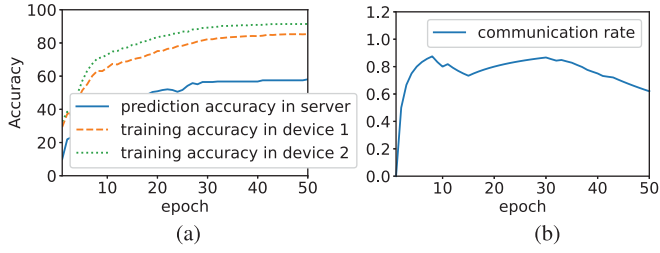


Fig. 5. The training and test performance of Algorithm 1 over CIFAR-10 dataset. (a) The accuracy in local training and central test. (b) The communication rate in the training process.

between the devices and the server has been reduced. Therefore, the proposed algorithm can save communication for more than 20% percentage while ensuring performance in model training for this experiment.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the even-triggered federated learning to reduce the communication overhead among a group of devices and a server. Based on stochastic gradient descent at the devices and model averaging at the server, an event-triggered federated learning algorithm was proposed. Under mild conditions, we established asymptotic convergence of the algorithm in expectation and almost surely. Moreover, the asymptotic normality and asymptotic efficiency of the estimation error were investigated. Regarding the future work, it is interesting to quantify the communication rate theoretically. In addition, we would like to extend the developed results to decentralized architectures under general graph conditions. Some practical communication problems, such as bandwidth constraint, channel gain and transmission delay, are also interesting for investigation. The co-employment of event-triggered communication and model sparsification/compression is also interesting to study.

APPENDIX

A. Proof of Proposition II.1

Let $\bar{w}(t) = \frac{1}{n} \sum_{j=1}^n w_j(t)$ be the average model of all n devices. It follows from Algorithm 1 that

$$\begin{aligned} \bar{w}(t) &= w_a(\tau_{k_a(t-1)}^a) \\ &\quad - \eta(t-1) \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a(t-1)}^a), \xi_j(t-1)) \right). \end{aligned} \quad (4)$$

It follows from Lemma A.1 that

$$\begin{aligned} &\mathbb{E}_{\xi(t)} \{f(\bar{w}(t))\} - f(w_a(\tau_{k_a(t-1)}^a)) \\ &\leq \mathbb{E}_{\xi(t)} \left\{ \nabla f^\top(w_a(\tau_{k_a(t-1)}^a)) (\bar{w}(t) - w_a(\tau_{k_a(t-1)}^a)) \right. \\ &\quad \left. + \frac{L_f}{2} \|\bar{w}(t) - w_a(\tau_{k_a(t-1)}^a)\|^2 \right\} \\ &\stackrel{(a)}{=} A + B, \end{aligned}$$

where (a) holds due to (4) and

$$\begin{aligned} A &= -\eta(t-1) \nabla f^\top(w_a(\tau_{k_a(t-1)}^a)) h_1(t) \\ h_1(t) &= \mathbb{E}_{\xi(t)} \left\{ \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a(t-1)}^a), \xi_j(t-1)) \right) \right\} \\ B &= \frac{\eta^2(t-1) L_f}{2} h_2(t) \\ h_2(t) &= \mathbb{E}_{\xi(t)} \left\{ \left\| \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a(t-1)}^a), \xi_j(t-1)) \right) \right\|^2 \right\}. \end{aligned}$$

Next, we consider the two terms A and B , respectively. Regarding A , it follows from Assumption II.2 1) that

$$\begin{aligned} A &= -\eta(t-1) \nabla f^\top(w_a(\tau_{k_a(t-1)}^a)) \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a(t-1)}^a)) \right) \\ &= -\eta(t-1) \left\| \nabla f(w_a(\tau_{k_a(t-1)}^a)) \right\|^2. \end{aligned}$$

Regarding B ,

$$\begin{aligned} B &\leq \frac{\eta^2(t-1) L_f}{2} \mathbb{E}_{\xi(t)} \left\{ 2 \left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(w_a(\tau_{k_a(t-1)}^a)) \right\|^2 \right. \\ &\quad \left. + 2 \left\| \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a(t-1)}^a), \xi_j(t-1)) \right. \right. \right. \\ &\quad \left. \left. \left. - \nabla f_j(w_a(\tau_{k_a(t-1)}^a)) \right) \right\|^2 \right\} \\ &\leq p_1 \eta^2(t-1) + L_f \eta^2(t-1) \left\| \nabla f(w_a(\tau_{k_a(t-1)}^a)) \right\|^2, \end{aligned}$$

where $p_1 = \frac{L_f q_1}{2}$ obtained from Assumption II.2 2).

Thus, it holds that

$$\begin{aligned} &\mathbb{E}_{\xi(t)} \{f(\bar{w}(t))\} - f(w_a(\tau_{k_a(t-1)}^a)) \\ &\leq -\eta(t-1) \left\| \nabla f(w_a(\tau_{k_a(t-1)}^a)) \right\|^2 + p_1 \eta^2(t-1) \\ &\quad + L_f \eta^2(t-1) \left\| \nabla f(w_a(\tau_{k_a(t-1)}^a)) \right\|^2. \end{aligned} \quad (5)$$

Let $e_a(t) = w_a(t) - w_a(\tau_{k_a(t)}^a)$ and $e_j(t) = w_j(t) - w_j(\tau_{k_j(t)}^j)$ denote the event-triggered errors at the server and device j , respectively, $j = 1, 2, \dots, n$. It can be seen that $e_a(t) = 0$ (resp. $e_j(t) = 0$) if the event of the server (resp. device j) is triggered. According to the event-triggered mechanisms in Algorithm 1, it follows that $\|e_a(t)\| \leq \mu_a(t)$ and $\|e_j(t)\| \leq \mu_j(t)$, for $j = 1, 2, \dots, n$. Note that

$$\begin{aligned} w_a(t-1) &= \frac{1}{n} \sum_{j=1}^n w_j(\tau_{k_j(t-1)}^j) \\ &= \frac{1}{n} \sum_{j=1}^n (w_j(t-1) - e_j(t-1)) \\ &= \bar{w}(t-1) - \frac{1}{n} \sum_{j=1}^n e_j(t-1), \end{aligned} \quad (6)$$

and

$$\begin{aligned} w_a(\tau_{k_a(t-1)}^a) &= w_a(t-1) + w_a(\tau_{k_a(t-1)}^a) - w_a(t-1) \\ &= w_a(t-1) - e_a(t-1) \\ &= \bar{w}(t-1) - e_0(t-1), \end{aligned} \quad (7)$$

where

$$e_0(t-1) = \frac{1}{n} \sum_{j=1}^n e_j(t-1) + e_a(t-1).$$

It holds that

$$\|e_0(t)\| \leq \frac{1}{n} \sum_{j=1}^n \mu_j(t) + \mu_a(t) \triangleq \mu_0(t). \quad (8)$$

It follows from (7) and Lemma A.1 that

$$\begin{aligned} f(w_a(\tau_{k_a(t-1)}^a)) &= f(\bar{w}(t-1) - e_0(t-1)) \\ &\leq f(\bar{w}(t-1)) - \nabla f^\top(\bar{w}(t-1))e_0(t-1) \\ &\quad + \frac{L_f}{2} \|e_0(t-1)\|^2. \end{aligned} \quad (9)$$

Taking expectation on both sides of the above inequality yields

$$\begin{aligned} \mathbb{E}\{f(w_a(\tau_{k_a(t-1)}^a))\} &\leq \mathbb{E}\{f(\bar{w}(t-1))\} \\ &\quad - \mathbb{E}\{\nabla f^\top(\bar{w}(t-1))e_0(t-1)\} + \frac{L_f}{2} \mathbb{E}\{\|e_0(t-1)\|^2\}. \end{aligned} \quad (10)$$

From the above inequality, (5) and $\eta^2(t) = o(\eta(t))$, there is a time T_0 , such that for any $t \geq T_0$, it follows that

$$\begin{aligned} \mathbb{E}\{f(\bar{w}(t))\} - \mathbb{E}\{f(\bar{w}(t-1))\} \\ \leq -\frac{\eta(t-1)}{2} \mathbb{E}\{\|\nabla f(w_a(\tau_{k_a(t-1)}^a))\|^2\} + p_1\eta^2(t-1) \\ - \mathbb{E}\{\nabla f^\top(\bar{w}(t-1))e_0(t-1)\} + \frac{L_f}{2} \mathbb{E}\{\|e_0(t-1)\|^2\}. \end{aligned} \quad (11)$$

It follows from (8) that

$$\begin{aligned} -\mathbb{E}\{\nabla f^\top(\bar{w}(t-1))e_0(t-1)\} &+ \frac{L_f}{2} \mathbb{E}\{\|e_0(t-1)\|^2\} \\ &\leq \mathbb{E}\{\|\nabla f^\top(\bar{w}(t-1))\| \|e_0(t-1)\|\} + \frac{L_f}{2} \mu_0^2(t-1) \\ &\leq \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|\} \mu_0(t-1) + \frac{L_f}{2} \mu_0^2(t-1) \\ &\stackrel{(a)}{\leq} \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\}^{1/2} \mu_0(t-1) + \frac{L_f}{2} \mu_0^2(t-1) \\ &\leq \left(\mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} + 1\right) \mu_0(t-1) + \frac{L_f}{2} \mu_0^2(t-1), \end{aligned} \quad (12)$$

where (a) follows from the Lyapunov inequality and the last inequality holds since $\mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\}$ is either larger than 1 or in $[0,1]$.

Note that

$$\begin{aligned} &-\mathbb{E}\{\|\nabla f(w_a(\tau_{k_a(t-1)}^a))\|^2\} \\ &= -\mathbb{E}\{\|\nabla f(w_a(\tau_{k_a(t-1)}^a)) - \nabla f(\bar{w}(t-1))\|^2\} \\ &\quad + \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} \end{aligned}$$

$$\begin{aligned} &\stackrel{(a)}{\leq} b_1 \mathbb{E}\{\| -e_0(t-1) \| \|\nabla f(\bar{w}(t-1))\| \\ &\quad - \|\nabla f(\bar{w}(t-1))\|^2\} \\ &\stackrel{(b)}{\leq} b_1 \mu_0(t-1)(1 + \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\}) \\ &\quad - \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\}, \end{aligned} \quad (13)$$

where $b_1 = \max\{\frac{2}{n} \sum_{j=1}^n L_i^1, 2\}$ and (a) follows from Assumption II.1 and $w_a(\tau_{k_a(t-1)}^a) - \bar{w}(t-1) = -e_0(t-1)$ in (7), and (b) holds since $\mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\}$ is either larger than 1 or in $[0,1]$.

Then it follows from (11)–(13) that

$$\begin{aligned} &\mathbb{E}\{f(\bar{w}(t))\} - \mathbb{E}\{f(\bar{w}(t-1))\} \\ &\leq \left(-\frac{\eta(t-1)}{2} + b_1 \mu_0(t-1) \frac{\eta(t-1)}{2} + \mu_0(t-1)\right) \\ &\quad \times \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} + b_1 \mu_0(t-1) \frac{\eta(t-1)}{2} \\ &\quad + p_1 \eta^2(t-1) + \mu_0(t-1) + \frac{L_f}{2} \mu_0^2(t-1). \end{aligned} \quad (14)$$

Under Assumption II.3, $\mu_0(t-1) \rightarrow 0$ and $\eta(t-1) \rightarrow 0$ as $t \rightarrow \infty$ and $\mu_0(t) = o(\eta(t))$, there exist a scalar b_2 and an integer $T_1 \geq T_0 > 0$, such that when $t \geq T_1$, it holds that

$$\begin{aligned} &\mathbb{E}\{f(\bar{w}(t))\} - \mathbb{E}\{f(\bar{w}(t-1))\} \\ &\leq -\frac{\eta(t-1)}{4} \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} \\ &\quad + b_2(\mu_0(t-1) + \eta^2(t-1)). \end{aligned} \quad (15)$$

Since $f(w^*)$ is the minimum loss function, $f(\bar{w}(t)) - f(w^*) \geq 0$. It follows from (15) that for $t \geq T_1$

$$\begin{aligned} &\mathbb{E}\{f(\bar{w}(t+1)) - f(w^*)\} - (\mathbb{E}\{f(\bar{w}(t)) - f(w^*)\}) \\ &\leq -\frac{\eta(t)}{4} \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} + b_2(\mu_0(t) + \eta^2(t)). \end{aligned}$$

Due to $\sum_{t=T_1}^{\infty} (\mu_0(t) + \eta^2(t)) < \infty$ and Lemma A.2, $\sum_{t=T_1}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} < \infty$, which together with $\sum_{t=0}^{T_1-1} \eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} < \infty$ leads to $\sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} < \infty$. To attain the conclusion, we derive

$$\begin{aligned} &\sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(w_a(t))\|^2\} \\ &= \sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(w_a(t)) - \nabla f(\bar{w}(t)) + \nabla f(\bar{w}(t))\|^2\} \\ &\stackrel{(a)}{\leq} 2b_3 \sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|w_a(t) - \bar{w}(t)\|^2\} \\ &\quad + 2 \sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} \\ &\stackrel{(b)}{\leq} 2b_3 \sum_{t=0}^{\infty} \eta(t) \mu_0^2(t) + 2 \sum_{t=0}^{\infty} \eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} < \infty, \end{aligned}$$

where $b_3 = (\frac{1}{n} \sum_{j=1}^n L_i^1)^2 > 0$, (a) is due to Assumption II.1 and the inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$, and (b) is from (6).

B. Proof of Theorem II.1

According to (15), there is a constant $b_2 > 0$ and an integer $T_1 > 0$, such that for any $t \geq T_1$, it holds that

$$\begin{aligned} & \mathbb{E}\{f(\bar{w}(t))\} - \mathbb{E}\{f(\bar{w}(t-1))\} \\ & \leq -\frac{\eta(t-1)}{4} \mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} \\ & \quad + b_2(\mu_0(t-1) + \eta^2(t-1)) \\ & \leq -\frac{\eta(t-1)s}{2} (\mathbb{E}\{f(\bar{w}(t-1))\} - f(w^*)) \\ & \quad + b_2(\mu_0(t-1) + \eta^2(t-1)), \end{aligned} \quad (16)$$

where the last inequality holds due to the following fact which is obtained according to Theorem 2.1.10 of [22] under the s -strong convexity condition in Assumption II.3:

$$\mathbb{E}\{\|\nabla f(\bar{w}(t-1))\|^2\} \geq 2s (\mathbb{E}\{f(\bar{w}(t-1))\} - f(w^*)).$$

From (16), it holds that

$$\begin{aligned} & \mathbb{E}\{f(\bar{w}(t))\} - f(w^*) \leq b_2(\mu_0(t-1) + \eta^2(t-1)) \\ & \quad + \left(1 - \frac{s\eta(t-1)}{2}\right) (\mathbb{E}\{f(\bar{w}(t-1))\} - f(w^*)). \end{aligned}$$

According to Lemma A.3, let $\alpha_t = s\eta(t)/2$ and $\beta_t = b_2(\mu_0(t) + \eta^2(t))$. Then under Assumption II.3, by noting that $\delta_0 = \rho_0 = \min\{1, \rho - 1\} \in (0, 1]$ and $\alpha_0 = 2\eta_0/s \geq 0$, we obtain that for any $\delta \in [0, \min\{\rho_0, s/(2\eta_0)\}]$,

$$\mathbb{E}\{f(\bar{w}(t))\} - f(w^*) = o(\eta^\delta(t)). \quad (17)$$

Recall from (6) that $w_a(t) = \bar{w}(t) - \frac{1}{n} \sum_{j=1}^n e_j(t)$, then it follows from Lemma A.1 that

$$\begin{aligned} f(w_a(t)) & \leq f(\bar{w}(t)) + \|\nabla f(\bar{w}(t))\| \left\| \frac{1}{n} \sum_{j=1}^n e_j(t) \right\| \\ & \quad + \frac{L_f}{2} \left\| \frac{1}{n} \sum_{j=1}^n e_j(t) \right\|^2. \end{aligned} \quad (18)$$

By taking the expectation and noting $\left\| \frac{1}{n} \sum_{j=1}^n e_j(t) \right\| \leq \mu_0(t)$, we derive

$$\begin{aligned} & \mathbb{E}\{f(w_a(t))\} - f(w^*) \leq \mathbb{E}\{f(\bar{w}(t))\} - f(w^*) \\ & \quad + \mu_0(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|\} + \frac{L_f}{2} \mu_0^2(t). \end{aligned} \quad (19)$$

It follows from Proposition II.1 that $\eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\} = o(1)$. Then

$$\begin{aligned} \mu_0(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|\} & = \frac{\mu_0(t)}{\eta^{\frac{1}{2}}(t)} \mathbb{E}\{\|\sqrt{\eta(t)} \nabla f(\bar{w}(t))\|\} \\ & \leq \frac{\mu_0(t)}{\eta^{\frac{1}{2}}(t)} \sqrt{\eta(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|^2\}} \\ & = o(\eta^{\rho-\frac{1}{2}}(t)), \end{aligned}$$

where the inequality is due to Hölder inequality.

By noting (17) and (19), from $\mu_0(t) \mathbb{E}\{\|\nabla f(\bar{w}(t))\|\} + \frac{L_f}{2} \mu_0^2(t) = o(\eta^{\rho-\frac{1}{2}}(t))$ and $\delta \leq \rho - 1 < \rho - \frac{1}{2}$, the error

$\mathbb{E}\{f(w_a(t))\} - f(w^*)$ decays to zero as fast as $\mathbb{E}\{f(\bar{w}(t))\} - f(w^*)$, which leads to the second conclusion. The first conclusion directly follows from the s -strong connectivity in Assumption II.3 and the second conclusion.

C. Proof of Theorem II.2

In the following, we prove the convergence on the iterates $w_a(t)$, which lead to the almost-sure convergence on $f(w_a(t))$ according to Lemma A.1.

From (4) and (7), it follows that

$$\begin{aligned} \bar{w}(t) & = \bar{w}(t-1) - e_0(t-1) \\ & \quad - \eta(t-1) \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(w_a(\tau_{k_a}^a(t-1))), \xi_j(t-1) \right). \end{aligned} \quad (20)$$

Recall from (6) that $w_a(t) = \bar{w}(t) - \frac{1}{n} \sum_{j=1}^n e_j(t)$. Substituting this equality into (20) yields

$$w_a(t) = w_a(t-1) - \eta(t-1) (\nabla f(w_a(t-1)) \quad (21)$$

$$+ \epsilon_1(t-1) + \epsilon_2(t-1)), \quad (22)$$

where

$$\begin{aligned} \epsilon_1(t) & = \frac{1}{n} \sum_{j=1}^n ((\nabla f_j(w_a(\tau_{k_a}^a(t))), \xi_j(t)) - \nabla f_j(w_a(\tau_{k_a}^a(t)))), \\ \epsilon_2(t) & = \frac{\hat{e}_0(t)}{\eta(t)} + \frac{1}{n} \sum_{j=1}^n \nabla f_j(w_a(\tau_{k_a}^a(t))) - \frac{1}{n} \sum_{j=1}^n \nabla f_j(w_a(t)), \end{aligned} \quad (23)$$

in which $\hat{e}_0(t-1) = e_0(t-1) + \frac{1}{n} \sum_{j=1}^n e_j(t) - \frac{1}{n} \sum_{j=1}^n e_j(t-1)$ subject to $\|\hat{e}_0(t-1)\| \leq 2\mu_0(t-1)$ due to the inequality $\left\| \frac{1}{n} \sum_{j=1}^n e_j(t) \right\| \leq \mu_0(t)$ and the monotonicity $\mu_0(t) \leq \mu_0(t-1)$.

Given the dynamics equation (21), in order to use the stochastic approximation results in Lemma A.5, besides the verification of almost-sure boundedness of $w_a(t)$, we need to verify C0–C3 to attain the first conclusion; and verify C0, C1'–C3' for the second conclusion.

The proof of the first conclusion: First, we verify C0. Let $x^0 = w^*$, $v(x) = f(x) - f(w^*)$, and $g(x) = -\nabla f(x)$. Since the optimal solution for the optimization problem (1) is unique, it holds that $g^\top(x) \nabla v(x) = -\nabla f^\top(x) \nabla f(x) = -\|\nabla f(x)\|^2 < 0$ for any $x \neq w^*$. Thus, C0 holds.

Second, we verify C1–C3 for the first conclusion. Let $\eta(t) = a_t$, then C1 holds. Regarding $\epsilon_2(t-1)$, it follows from (7) and Assumption II.1 that

$$\begin{aligned} \|\epsilon_2(t-1)\| & \leq \frac{2\mu_0(t-1)}{\eta(t-1)} + L_m \left\| w_a(\tau_{k_a}^a(t-1)) - w_a(t-1) \right\| \\ & \leq \bar{L} \frac{\mu_0(t-1)}{\eta(t-1)}, \end{aligned} \quad (24)$$

where $L_m = \max\{L_i\}_{i=1}^n$ and $\bar{L} > 0$ is a constant. Regarding $\epsilon_1(t)$, it follows from Assumption II.2 2) that there is a constant $q_0 > 0$, such that

$$\sup_t \mathbb{E}\{\|\epsilon_1(t)\|^2 | \mathcal{F}(t-1)\} < q_0, \quad \text{a. s.} \quad (25)$$

In addition, due to Assumption II.2 1) it holds that $\mathbb{E}\{\epsilon_1(t)|\mathcal{F}(t-1)\} = 0$, a.s.. Using $\mathbb{E}\{\mathbb{E}\{X|\mathcal{F}(t-1)\} = \mathbb{E}\{X\}$ and Hölder's inequality on (25) leads to $\mathbb{E}\{\|\epsilon_1(t)\|\} < \infty$. Thus, $\{\epsilon_1(t), \mathcal{F}(t)\}$ is a martingale difference sequence (MDS). Due to $\sum_{t=0}^{\infty} \eta^2(t) < \infty$ in Assumption II.3 and [23], Theorem B.2.1], it holds that $\sum_{t=1}^{\infty} \eta(t)\epsilon_1(t) < \infty$, a.s.. Moreover, $\sum_{t=1}^{\infty} \eta(t)\|\epsilon_2(t)\| \leq \bar{L} \sum_{t=1}^{\infty} \mu_0(t) < \infty$. Then it follows from that $\sum_{t=1}^{\infty} \eta(t)(\epsilon_1(t) + \epsilon_2(t)) < \infty$, a.s., meaning C2 holds. Since $g(x) = -\nabla f(x)$ is a continuous function, it is measurable. It follows from Assumption II.1 that $g(x)$ is locally bounded. Then C3 holds.

For the first conclusion, it remains to prove the boundedness of $w_a(t)$ by employing Lemma A.4. Let $\eta(t) = a_t$, the first condition in Lemma A.4 holds under Assumption II.3. The second and third conditions hold from the above analysis and from Assumption II.1, respectively.

The proof of the second conclusion: We verify C1'–C3' for the second conclusion. Under Assumption II.4, C1' holds. Let $\epsilon'_t = \epsilon_1(t)$ and $\epsilon''_t = \epsilon_2(t)$. Similar to the verification for C2, it follows from Assumption II.4 that $\sum_{t=0}^{\infty} \eta^{(1-\delta)}(t)\epsilon_1(t) < \infty$, a.s., where $\delta \in (0, \frac{1}{2})$ is the one in Assumption II.4. In addition, it follows from Assumption II.4 and (24) that $\epsilon_2(t) = O(\eta^\delta(t))$, a.s.. Thus, C2' holds. Let $g(w) = -\nabla f(w)$ and $F = -\nabla^2 f(w^*)$. Then we obtain a first-order vector-valued Taylor approximation of $g(w)$ according to [24, Theorem 5]. Under Assumption II.4 3), it follows from [22, Theorem 2.1.11] that $\nabla^2 f(w^*) \geq sI$. Then C3' holds due to $\nabla^2 f(w^*) \geq sI > \eta_0 \delta I$. The assumptions for the second conclusion also ensure the boundedness of the iteration, since all the conditions in Lemma A.4 are fulfilled. Thus, the second conclusion is attained.

D. Proof of Theorems II.3

In order to use the stochastic results in Lemma A.5, we aim to verify C0, C1, C2, and C3 which ensure the asymptotic convergence of iteration, as well as verify C2'' and C3''. Under the conditions of this theorem, C0 is fulfilled like the analysis in the proof of Theorem II.2. Due to $a_t = \eta(t) = \frac{1}{t^v}$ with $v \in (\frac{2}{3}, 1)$, C1 is fulfilled. Let $g(w) = -\nabla f(w)$ and $F = -\nabla^2 f(w^*)$. Similar to the proof in Theorem II.2, it follows from Assumption II.5 2) and the second-order vector-valued Taylor approximation of $g(w)$ [24, Theorem 5] that C3'' holds. Since C3'' is a sufficient condition of C3, C3 holds. It remains to verify C2 and C2''. The verification of C2 is similar to the one in the proof of Theorem II.2 by noting that such a setting $\eta(t) = \frac{1}{t^v}$ with $v \in (\frac{2}{3}, 1)$ and $\mu_0(t) = o(\eta(t)^{\frac{3}{2}})$ ensures $\sum_{t=0}^{\infty} \eta^2(t) < \infty$ and $\sum_{t=1}^{\infty} \mu_0(t) < \infty$. Thus, C2 is satisfied. Similar to the analysis in the proof of Theorem II.2, $\epsilon_2(t) = o(\sqrt{\eta(t)})$ ensures (36). From the proof of Theorem II.2, $\{\epsilon_1(t), \mathcal{F}(t)\}$ is an MDS satisfying the first two conditions in (37). Recall from (23) that

$$\begin{aligned} \epsilon_1(t) &= \frac{1}{n} \sum_{j=1}^n ((\nabla f_j(w_a(\tau_{k_a}^a(t))), \xi_j(t+1)) \\ &\quad - \nabla f_j(w_a(\tau_{k_a}^a(t)))) = \frac{1}{n} \sum_{j=1}^n \bar{\epsilon}_j(t). \end{aligned}$$

Then it follows from (2) that

$$\lim_{t \rightarrow \infty} \mathbb{E}\{\epsilon_1(t)\epsilon_1^T(t)|\mathcal{F}(t-1)\} = \frac{1}{n^2} \sum_{j=1}^n S_j = S_0.$$

According to the Dominated Convergence Theorem [23], P331], it holds that

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{E}\{\epsilon_1(t)\epsilon_1^T(t)\} &= \lim_{t \rightarrow \infty} \mathbb{E}\{\mathbb{E}\{\epsilon_1(t)\epsilon_1^T(t)|\mathcal{F}(t-1)\}\} \\ &= \mathbb{E}\{\lim_{t \rightarrow \infty} \mathbb{E}\{\epsilon_1(t)\epsilon_1^T(t)|\mathcal{F}(t-1)\}\} = S_0. \end{aligned}$$

Note that $\epsilon_1(t) = \frac{1}{n} \sum_{j=1}^n \bar{\epsilon}_j(t)$, then it follows from (2) that $\sup_t \mathbb{E}\{\|\epsilon_1(t)\|^p\} < \infty$. Due to $p > 2$, it follows from [25, Corollary 6.6] that $\|\epsilon_1(t)\|^2$ is uniformly differentiable, i.e., $\lim_{a \rightarrow \infty} \sup_t \mathbb{E}\{\|\epsilon_1(t)\|^2 I_{\|\epsilon_1(t)\| > a}\} \rightarrow 0$, meaning C2'' holds. In addition, similar to the analysis for Theorem II.2, the boundedness of iterates $w_a(t)$ is guaranteed under the conditions. Thus, all the conditions are fulfilled. The conclusions on convergence in distribution follow from (c) of Lemma A.5.

To prove (3), we use Lemma A.1 to obtain $\mathbb{E}\{f(\hat{w}_a(t))\} - f(w^*) \leq \frac{L_f}{2} \mathbb{E}\{\|\hat{w}_a(t) - w^*\|^2\}$. Due to $\sqrt{t}(\hat{w}_a(t) - w^*) \xrightarrow[t \rightarrow \infty]{d} N(0, \bar{S})$ and $\bar{S} = H^{-1}(\frac{1}{n^2} \sum_{j=1}^n S_j)(H^{-1})^T \leq \frac{1}{n} S_0$, where $S_0 = H^{-1} \max_{j=1}^n S_j (H^{-1})^T$, it holds that $\mathbb{E}\{\|\hat{w}_a(t) - w^*\|^2\} = O(\frac{1}{nt})$.

E. Useful Lemmas

Lemma A.1: ([22], Theorem 2.1.5) Under Assumption II.1, for any $w_a, w_b \in \mathbb{R}^m$, it holds that

$$f(w_b) \leq f(w_a) + \nabla f^T(w_a)(w_b - w_a) + \frac{L_f}{2} \|w_b - w_a\|^2,$$

where $L_f = \max_{i=1,2,\dots,n} \{L_i\}$.

Lemma A.2: ([26]) Let $\{v(t)\}$ be a non-negative sequence, such that for any $t \geq 0$, it holds that

$$v(t+1) \leq (1+a(t))v(t) - u(t) + w(t),$$

where $a(t) \geq 0, u(t) \geq 0$ and $w(t) \geq 0$ with $\sum_{t=0}^{\infty} a(t) < \infty$ and $\sum_{t=0}^{\infty} w(t) < \infty$. Then the sequence $\{v(t)\}$ converges to $v \geq 0$ and $\sum_{t=0}^{\infty} u(t) < \infty$.

Lemma A.3: Assume that $\alpha_t > 0, \alpha_t \rightarrow 0, \sum_{t=1}^{\infty} \alpha_t = \infty$, and

$$\lim_{t \rightarrow \infty} \left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} \right) = \alpha_0 \geq 0. \quad (26)$$

Consider the iteration $x_{t+1} = (1 - \alpha_t)x_t + \beta_t$, where $\beta_t = O(\alpha_t^{1+\delta_0})$ with $0 < \delta_0 \leq 1$, then for any $\delta \in [0, \min\{\delta_0, 1/\alpha_0\})$, it holds that

$$x_t = o(\alpha_t^\delta).$$

Proof: Let $z_t = \frac{x_t}{\alpha_t^\delta}$, then it holds that

$$z_{t+1} = (1 - \alpha_t) \left(\frac{\alpha_t}{\alpha_{t+1}} \right)^\delta z_t + \frac{\beta_t}{\alpha_{t+1}^\delta}. \quad (27)$$

We seek to prove $z_t \xrightarrow[t \rightarrow \infty]{} 0$.

It follows from (26) that

$$\begin{aligned} \frac{\alpha_t}{\alpha_{t+1}} &= \frac{\alpha_t}{\alpha_{t+1}} - 1 + 1 \\ &= 1 + \alpha_t \left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} \right) \\ &= 1 + \alpha_0 \alpha_t + o(\alpha_t). \end{aligned} \quad (28)$$

Due to $\alpha_t > 0$ and $\alpha_t \rightarrow 0$, there exists a time $t_0 > 0$, such that $\alpha_t \in (0, 1)$ and $\alpha_0 \alpha_t + o(\alpha_t) > -1$ for $t \geq t_0$. Then for $t \geq t_0$, we derive

$$\begin{aligned} (1 - \alpha_t) \left(\frac{\alpha_t}{\alpha_{t+1}} \right)^\delta &\stackrel{(a)}{\leq} (1 - \alpha_t)(1 + \delta(\alpha_0 \alpha_t + o(\alpha_t))) \\ &= (1 - (1 - \alpha_0 \delta) \alpha_t + o(\alpha_t)) \\ &\stackrel{(b)}{\leq} \exp(-(1 - \alpha_0 \delta) \alpha_t + o(\alpha_t)), \end{aligned} \quad (29)$$

where (a) follows from (28) and $(1 + x)^r \leq 1 + rx$ for $0 \leq r \leq 1$ and $x > -1$, and (b) is due to $1 + x \leq \exp(x)$ for any $x \in \mathbb{R}$.

It follows from (27)–(29) that for $t \geq t_0$,

$$\begin{aligned} z_{t+1} &\leq \exp(-(1 - \alpha_0 \delta) \alpha_t + o(\alpha_t)) z_t + \frac{\alpha_t^\delta}{\alpha_{t+1}^\delta} \frac{\beta_t}{\alpha_t^\delta} \\ &\leq \exp(-(1 - \alpha_0 \delta) \alpha_t + o(\alpha_t)) z_t \\ &\quad + (1 + \alpha_0 \alpha_t + o(\alpha_t))^\delta \frac{\beta_t}{\alpha_t^\delta}. \end{aligned} \quad (30)$$

Due to $\delta < \delta_0$ and $\alpha_t \rightarrow 0$, it follows from $\beta_t = O(\alpha_t^{1+\delta_0})$ that $\frac{\beta_t}{\alpha_t^\delta} = o(\alpha_t)$, then for any $\epsilon > 0$, there is a time $t_1 \geq t_0$ such that for $t \geq t_1$, it holds that

$$\frac{\beta_t}{\alpha_t^\delta} \leq \epsilon \alpha_t. \quad (31)$$

Let $c_1 = \frac{(1-\alpha_0\delta)}{2}$, which is positive due to $\delta < \min\{\delta_0, \frac{1}{\alpha_0}\}$. Then for large enough t_2 and $t > t_2$, we have that

$$\begin{aligned} \alpha_t &\leq \alpha_t + (\alpha_t - c_1 \alpha_t^2) \\ &= 2 \left(\alpha_t - \frac{c_1 \alpha_t^2}{2} \right) \\ &\leq \frac{2}{c_1} (1 - \exp(-c_1 \alpha_t)), \end{aligned} \quad (32)$$

where the last inequality is due to $e^{-x} \leq 1 - x + x^2/2$, for $\forall x \geq 0$.

There is a time $t_3 \geq \max\{t_1, t_2\}$, such that for $t \geq t_3$, it holds that $o(\alpha_t) \leq \frac{(1-\alpha_0\delta)\alpha_t}{2}$ and $1 + \alpha_0 \alpha_t + o(\alpha_t) \leq 2$. For ease of notations, let $\sum_{i=j}^t \alpha_i = 0$ if $j > t$. It follows from (30) that for $t_3 \geq \max\{t_1, t_2\}$

$$\begin{aligned} z_{t+1} &\leq \exp(-c_1 \alpha_t) z_t + 2^\delta \frac{\beta_t}{\alpha_t^\delta} \leq \exp \left(-c_1 \sum_{j=t_3}^t \alpha_j \right) z_{t_3} \\ &\quad + 2^\delta \sum_{j=t_3}^t \exp \left(-c_1 \sum_{i=j+1}^t \alpha_i \right) \frac{\beta_j}{\alpha_j^\delta}. \end{aligned}$$

It follows from $\sum_{t=1}^\infty \alpha_t = \infty$ that $\exp(-c_1 \sum_{j=t_3}^t \alpha_j) z_{t_3} \rightarrow 0$ as $t \rightarrow \infty$. Next, we prove that the second term goes to zero as well.

For $t \geq t_3$, it holds that

$$\begin{aligned} &\sum_{j=t_3}^t \exp \left(-c_1 \sum_{i=j+1}^t \alpha_i \right) \frac{\beta_j}{\alpha_j^\delta} \\ &\stackrel{(c)}{\leq} \epsilon \sum_{j=t_3}^t \exp \left(-c_1 \sum_{i=j+1}^t \alpha_i \right) \alpha_j \\ &\stackrel{(d)}{\leq} \frac{2\epsilon}{c_1} \sum_{j=t_3}^t \exp \left(-c_1 \sum_{i=j+1}^t \alpha_i \right) (1 - \exp(-c_1 \alpha_j)) \\ &= \frac{2\epsilon}{c_1} \sum_{j=t_3}^t \left(\exp \left(-c_1 \sum_{i=j+1}^t \alpha_i \right) - \exp \left(-c_1 \sum_{i=j}^t \alpha_i \right) \right) \\ &\leq \frac{2\epsilon}{c_1}, \end{aligned}$$

where (c) and (d) are due to (31) and (32), respectively. Since $\epsilon > 0$ is arbitrarily small, the term $\sum_{j=t_3}^t \exp(-c_1 \sum_{i=j+1}^t \alpha_i) \frac{\beta_j}{\alpha_j^\delta}$ tends to zero as t goes to infinity. \square

To study the convergence of the following iteration

$$x_{t+1} = x_t + a_t(g(x_t) + \epsilon_t), \quad (33)$$

we introduce some conditions:

C0 A continuously differentiable function $v(\cdot): \mathbb{R}^l \rightarrow \mathbb{R}$ exists such that

$$g^\top(x) \nabla v(x) < 0, \quad \forall x \neq x^0.$$

C1 $a_t > 0$, $a_t \rightarrow 0$, $\sum_{t=1}^\infty a_t = \infty$.

C1' In addition to C1, a_t satisfies

$$\frac{a_t - a_{t+1}}{a_t a_{t+1}} \rightarrow \alpha \geq 0. \quad (34)$$

C2 The noise sequence $\sum_{t=1}^\infty a_t \epsilon_t < \infty$, a.s..

C2' The noise sequence ϵ_t in (33) can be decomposed into two parts $\epsilon_t = \epsilon'_t + \epsilon''_t$ such that

$$\sum_{t=1}^\infty a_t^{1-\delta} \epsilon'_t < \infty, \quad \epsilon''_t = O(a_t^\delta), \quad \text{a.s.}, \quad (35)$$

for some $\delta \in (0, 1]$.

C2'' The noise ϵ_t in (33) can be decomposed into two parts $\epsilon_t = \epsilon'_t + \epsilon''_t$ such that

$$\epsilon''_t = o(\sqrt{a_t}), \quad \text{a.s.}, \quad (36)$$

and $\{\epsilon'_t, \mathcal{F}_t\}$ with $\mathcal{F}_t = \sigma\{x_0, \epsilon_i\}_{i=0}^t$ being an MDS satisfying the following conditions:

$$\begin{aligned} \mathbb{E}\{\epsilon'_t | \mathcal{F}_{t-1}\} &= 0, \quad \sup_t \mathbb{E}\{\|\epsilon'_t\|^2 | \mathcal{F}_{t-1}\} \leq \sigma \\ \lim_{t \rightarrow \infty} \mathbb{E}\{\epsilon'_t (\epsilon'_t)^\top | \mathcal{F}_{t-1}\} &= \lim_{t \rightarrow \infty} \mathbb{E}\{\epsilon'_t (\epsilon'_t)^\top\} = S_0, \\ \lim_{a \rightarrow \infty} \sup_t \mathbb{E}\{\|\epsilon'_t\|^2 I_{\|\epsilon'_t\| > a}\} &= 0, \quad \text{a.s.}, \end{aligned} \quad (37)$$

where $\sigma \geq 0$ is a constant.

C3 $g(\cdot)$ is measurable and locally bounded.

C3' $g(\cdot)$ is measurable and locally bounded, and is differentiable at x^0 such that as $x \rightarrow x^0$

$$\begin{aligned} g(x) &= F(x - x^0) + \tau(x), \quad \tau(x^0) = 0, \\ \tau(x) &= o(\|x - x^0\|). \end{aligned} \quad (38)$$

Both matrices F and $F + \alpha \delta I$ are stable, where α and δ are given in (34) and (35), respectively.

C3'' $g(\cdot)$ is measurable and locally bounded, and is differentiable at x^0 such that as $x \rightarrow x^0$

$$\|g(x) - F(x - x^0)\| \leq c \|x - x^0\|^2,$$

where $c > 0$ and F is stable.

Lemma A.4: The iterate x_t is bounded a.s. if all the following conditions hold:

1. $a_t > 0$, $\sum_{t=1}^{\infty} a_t = \infty$ and $\sum_{t=1}^{\infty} a_t^2 < \infty$.
2. The noise ϵ_t in (33) can be decomposed into two parts $\epsilon_t = \epsilon'_t + \epsilon''_t$, such that $\epsilon'_t = o(1)$, a.s., and $\{\epsilon'_t, \mathcal{F}_t\}$ with $\mathcal{F}_t = \sigma\{x_0, \epsilon_i\}_{i=0}^t$ is an MDS satisfying the following conditions:

$$\mathbb{E}\{\epsilon'_t | \mathcal{F}_{t-1}\} = 0, \quad \sup_t \mathbb{E}\{\|\epsilon'_t\|^2 | \mathcal{F}_{t-1}\} \leq \sigma, \quad (39)$$

where $\sigma \geq 0$ is a constant.

3. There exists a continuously differentiable function $f(x) > -\infty$, such that $g(\cdot) = -\nabla f(\cdot)$. In addition, $f(x)$ is radially unbounded (i.e., $f(x) \xrightarrow{x \rightarrow \infty} \infty$) and its gradient $\nabla f(\cdot)$ is Lipschitz continuous.

Proof: The proof follows from [27, Theorem 4.7.1] by verifying assumptions A4.1.2, A4.1.3 and A4.7.1–A4.7.3. To verify A4.7.2, it follows from [27, Example 6] by considering the first two conditions here. The rest of the assumptions can be directly verified under the provided conditions. \square

Lemma A.5: [15], [23] Consider the iteration (33). Assume x_t is bounded almost surely. Then the following conclusions hold:

1. Let C0, C1, C2, and C3 hold. Then

$$\lim_{t \rightarrow \infty} x_t = x^0, \quad \text{a.s.}$$

2. Let C0, C1', C2', and C3' hold. Then

$$\|x_t - x^0\| = o(a_t^\delta), \quad \text{a.s.},$$

where δ is the one given in C2'.

3. Let $a_t = \frac{1}{t^v}$ with $v \in (\frac{2}{3}, 1)$. If $\lim_{t \rightarrow \infty} x_t = x^0$, a.s. and C0, C2'', and C3'' hold, then the following conclusions hold:

- (1) $\frac{1}{\sqrt{a_t}}(x_t - x^0)$ is asymptotically normal:

$$\frac{1}{\sqrt{a_t}}(x_t - x^0) \xrightarrow[t \rightarrow \infty]{d} N(0, S),$$

where $S = \int_0^\infty e^{Ft} S_0 e^{F^\top t} dt$ and F is the one given in C3''.

- (2) $\bar{x}_t := \frac{1}{t} \sum_{i=1}^t x_i - x^0$ is asymptotically efficient:

$$\sqrt{t} \bar{x}_t \xrightarrow[t \rightarrow \infty]{d} N(0, \bar{S}),$$

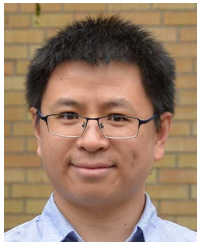
where $\bar{S} = F^{-1} S_0 (F^{-1})^\top$ and F is the one given in C3''.

Remark A.1: In Lemma A.5, from [23, Theorem 2.4.1] and the almost-sure boundedness of x_t , C2 satisfies [23, A2.2.3], then assertion 1) holds according to [23, Theorem 2.2.1]. Assertion 2) holds following [23, Theorem 3.1.1]. Assertions 3) is from [15, Lemma A.4].

REFERENCES

- [1] J. Xu, W. Du, Y. Jin, W. He, and R. Cheng, "Ternary compression for communication-efficient federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1162–1176, Mar. 2022.
- [2] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.
- [3] D. Rothchild et al., "FetchSGD: Communication-efficient federated learning with sketching," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria: PMLR, 2020, pp. 8253–8265.
- [4] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Statist.*, Palermo, Italy: PMLR, 2020, pp. 2021–2031.
- [5] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5986–5994, Jul. 2020.
- [6] Z. Yang et al., "Delay minimization for federated learning over wireless communication networks," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1–7.
- [7] C. Liu, T. J. Chua, and J. Zhao, "Time minimization in hierarchical federated learning," in *Proc. IEEE/ACM 7th Symp. Edge Comput. (SEC)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 96–106.
- [8] R. Chen, D. Shi, X. Qin, D. Liu, M. Pan, and S. Cui, "Service delay minimization for federated learning over mobile devices," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 990–1006, Apr. 2023.
- [9] X. Zhou, J. Zhao, H. Han, and C. Guet, "Joint optimization of energy consumption and completion time in federated learning," in *Proc. IEEE 42nd Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1005–1017.
- [10] J. Ren and J. Haupt, "A provably communication-efficient asynchronous distributed inference method for convex and nonconvex problems," *IEEE Trans. Signal Process.*, vol. 68, pp. 3325–3340, 2020.
- [11] J. Hamer, M. Mohri, and A. T. Suresh, "FedBoost: A communication-efficient algorithm for federated learning," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria: PMLR, 2020, pp. 3973–3983.
- [12] W. Li, Z. Wu, T. Chen, L. Li, and Q. Ling, "Communication-censored distributed stochastic gradient descent," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6831–6843, Nov. 2022.
- [13] T. Chen, Y. Sun, and W. Yin, "Communication-adaptive stochastic gradient methods for distributed learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 4637–4651, 2021.
- [14] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "FedPD: A federated learning framework with adaptivity to non-iid data," *IEEE Trans. Signal Process.*, vol. 69, pp. 6055–6070, 2021.
- [15] J. Lei, H.-F. Chen, and H.-T. Fang, "Asymptotic properties of primal-dual algorithm for distributed stochastic optimization over random networks with imperfect communications," *SIAM J. Control Optim.*, vol. 56, no. 3, pp. 2159–2188, 2018.
- [16] T. Yang et al., "A survey of distributed optimization," *Annu. Rev. Control*, vol. 47, pp. 278–305, May 2019.
- [17] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization," *IEEE Trans. Autom. Control*, vol. 68, no. 2, pp. 721–736, Feb. 2023.
- [18] X. Yi, L. Yao, T. Yang, J. George, and K. H. Johansson, "Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication," in *Proc. IEEE Conf. Decis. Control*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 3397–3402.
- [19] J. George and P. Gurram, "Distributed stochastic gradient descent with event-triggered communication," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 7169–7178.

- [20] X. Cao and T. Başar, “Decentralized online convex optimization with event-triggered communications,” *IEEE Trans. Signal Process.*, vol. 69, pp. 284–299, 2021.
- [21] S. Ghosh, B. Aquino, and V. Gupta, “EventGraD: Event-triggered communication in parallel machine learning,” *Neurocomputing*, vol. 483, pp. 474–487, Apr. 2022.
- [22] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137. Berlin, Germany: Springer-Verlag, 2018.
- [23] H.-F. Chen, *Stochastic Approximation and Its Applications*. Boston, MA, USA: Kluwer, 2002.
- [24] J. E. Chacón and T. Duong, *Multivariate Kernel Smoothing and Its Applications*. Boca Raton, FL, USA: Chapman and Hall/CRC Press, 2018.
- [25] A. S. Poznyak, *Advanced Mathematical Tools for Automatic Control Engineers: Stochastic Techniques*. Amsterdam, The Netherlands: Elsevier, 2009.
- [26] H. Robbins and D. Siegmund, “A convergence theorem for non negative almost supermartingales and some applications,” in *Optimizing Methods in Statistics*. New York, NY, USA: Elsevier, 1971, pp. 233–257.
- [27] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York, NY, USA: Springer-Verlag, 1978.



Xingkang He received the B.S. degree from the School of Mathematics at Hefei University of Technology in 2013 and the Ph.D. degree from the University of Chinese Academy of Sciences in 2018. He is now working with Ericsson, AB. Before this position, he worked as a Postdoctoral Research Associate with the Department of Electrical Engineering, University of Notre Dame from 2021 to 2022 and as a Postdoctoral Researcher with the Division of Decision and Control Systems, KTH Royal Institute of Technology from 2018 to 2021.

His research interests include security of cyber-physical systems, estimation and control of networked systems, and social networks. He was a recipient of best paper award in 2018 *IEEE Data Driven Control and Learning Systems Conference*.



Xinlei Yi received the B.S. and M.S. degrees in mathematics from the China University of Geoscience and Fudan University in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology in 2020. Currently, he is pursuing the Post-Doc with the same university. His current research interests include online optimization, distributed optimization, and event-triggered control.



Yanlong Zhao received the B.S. degree in mathematics from Shandong University, Jinan, China, in 2002 and the Ph.D. degree in systems theory from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2007. Since 2007, he has been with the AMSS, CAS, where he is currently a full Professor. His research interests include identification and control of quantized systems, information theory, and modeling of financial systems. He has been a Deputy Editor-in-Chief of *Journal of Systems*

and *Science and Complexity* and an Associate Editor of *Automatica*, *SIAM Journal on Control and Optimization*, *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS*, and *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS*. He served as a Vice-President of the Asian Control Association. He is now a Vice General Secretary of Chinese Association of Automation (CAA), a Vice-Chair of Technical Committee on Control Theory (TCCT), CAA, a Vice-President of IEEE CSS Beijing Chapter, and a Member of IFAC TC Modelling, Identification & Signal Processing.



Karl Henrik Johansson (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. He is a Professor with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology. His research interests are in networked control systems, cyber-physical systems, and applications in transportation, energy, and automation. He has served on the IEEE Control Systems Society Board of Governors, the

IFAC Executive Board, and the European Control Association Council. He has received several best paper awards and other distinctions from IEEE and ACM. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He has received the Future Research Leader Award from the Swedish Foundation for Strategic Research and the triennial Young Author Prize from IFAC. He is a fellow with the Royal Swedish Academy of Engineering Sciences, and he is an IEEE Distinguished Lecturer.



Vijay Gupta (Fellow, IEEE) received the B.Tech. degree from the Indian Institute of Technology, Delhi, and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology. He is working with the Elmore Family School of Electrical and Computer Engineering at Purdue University, having joined the faculty in May 2022. He received the 2018 Antonio J Rubert Award from the IEEE Control Systems Society, the 2013 Donald P. Eckman Award from the American Automatic Control Council, and a 2009 National Science

Foundation (NSF) CAREER Award. His research and teaching interests are broadly in the interface of communication, control, distributed computation, and human decision making.