This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

Differential Privacy Meets Federated Learning under Communication Constraints

Nima Mohammadi, Jianan Bai, Qiang Fan, Yifei Song, Yang Yi, and Lingjia Liu

***Abstract*—The performance of federated learning systems is bottlenecked by communication costs and training variance. The communication overhead problem is usually addressed by three communication-reduction techniques, namely, model compression, partial device participation, and periodic aggregation, at the cost of increased training variance. Different from traditional distributed learning systems, federated learning suffers from data heterogeneity (since the devices sample their data from possibly different distributions), which induces additional variance among devices during training. Various variance-reduced training algorithms have been introduced to combat the effects of data heterogeneity, while they usually cost additional communication resources to deliver necessary control information. Additionally, data privacy remains a critical issue in FL, and thus there have been attempts at bringing Differential Privacy to this framework as a mediator between utility and privacy requirements. This paper investigates the trade-offs between communication costs and training variance under a resource-constrained federated system theoretically and experimentally, and studies how communication reduction techniques interplay in a differentially private setting. The results provide important insights into designing practical privacy-aware federated learning systems.**

***Index Terms*—Federated Learning, Differential Privacy, Artificial Intelligence, Communication Constraints, and Training Variance**

I. Introduction

Artificial intelligence is expected to have a significant impact on Beyond 5G and 6G networks [1]. This is especially true for Internet of Things (IoT) [2] where massive connectivity is expected from heterogeneous devices. With the ever increasing importance of data privacy of these devices, federated learning emerges as a promising machine learning framework that enables the training of a shared model among multiple end devices and a parameter server without exchanging local data [3]–[6]. Assuming a total of N local devices and each device $i \in \{1, 2, \cdots, N\}$ possesses a local dataset $\mathcal{D}_i$ with $|\mathcal{D}_i|$ samples, the local objective of device i can be formulated as the following risk minimization problem

$$\min_{\mathbf{x}} f_i(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}_i} \ell(\mathbf{x}; \xi),$$

where $\mathbf{x} \in \mathbb{R}^d$ denotes the model parameter, $\xi \in \mathbb{R}^u$ is a training sample, and $\ell : \mathbb{R}^d \times \mathbb{R}^u \to \mathbb{R}$ is the sample-wise loss function. On the other hand, the goal of the parameter server is to find a single global model that would work well on the whole dataset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_N$. Accordingly, the global objective function can be represented as $f(\mathbf{x}) = \sum_{i=1}^{N} w_i f_i(\mathbf{x})$, where $w_i = |\mathcal{D}_i|/|\mathcal{D}|$ is the device weight proportional to the sample size [7].

Despite the privacy advantages of federated learning, which, as we discuss later, are prone to some challenges, its real-world implementation raises some issues regarding communication overhead and training variance:

A. Communication Overhead

Since the parameter server has no access to the local datasets, it needs to collect the local model updates from the local devices periodically, then send the aggregated model back to all devices. The communication rounds between the parameter server and the local devices result in a substantial communication overhead, especially when the number of devices is large. Moreover, the underlying communications between the parameter server and the local devices are usually imperfect and with limited capacity, and hence, it is imperative to design reliable and efficient federated learning algorithms under communication budget constraints.

From a communication point of view, federated learning systems realize model aggregation (uplink) through the multiple access channel (MAC) [8], and model distribution (downlink) through the broadcast channel (BC) [9]. Since a substantial number of local devices will send potentially different local model updates to the parameter server during model aggregation, the limited capacity of the uplink communications is usually the bottleneck of the system. Therefore, the existing literature suggests the following communication-reduction strategies to be utilized for enhancing the communication efficiency of federated learning systems, especially for the uplink (MAC) part:

1) Model Compression: The transmission of each full-accuracy single-precision floating-point value requires 32 bits. Since the model size is generally large in machine learning systems, transmitting model parameters with full accuracy can be prohibitively expensive. In this regard, a common strategy is to quantize the local model updates with some low-accuracy compressors or send only some important local parameters. Overall, the communication overhead can be significantly reduced via compression.

2) Partial Participation: Due to the straggler's effect (some devices becoming non-responding and inactive), the response time for some devices can be prohibitively long. Therefore, awaiting model updates from all devices is not desirable in practice. Furthermore, the capacity of the underlying MAC channel is usually limited and does not linearly increase with

The authors are with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060 USA. The work is supported in part by US National Science Foundation (NSF) under grant CCF-1937487. The corresponding author is L. Liu (ljliu@ieee.org).

the number of transmitting devices. A natural solution to resolve these issues is partial participation through scheduling: only $M < N$ devices will be scheduled to transmit during each communication slot (round).

3) Periodic Aggregation: The aggregation of local models requires synchronization among all active devices. Aggregating in each iteration of training, as in traditional distributed learning, results in a large communication overhead. Therefore, frequent model synchronization may become unrealistic and consume a lot of system resources for communication. A widely adopted strategy is to conduct several local iterations before synchronizing with the parameter server to save communication overhead. However, since all devices perform local updates in an unsynchronized way, the update direction can deviate from the global gradient direction, especially for non-i.i.d. settings.

It is important to note that the three communication-reduction strategies are inherently coupled under communication constraints. For example, the payload of the model parameters will inevitably impact how many devices (M) can be active under a fixed capacity of the MAC channel. Meanwhile, there is a clear trade-off between the payload of the model parameters and the period of the model aggregation under a fixed MAC capacity constraint: a larger payload will lead to a less frequent model aggregation. Therefore, a joint analysis seems necessary to provide a comprehensive analysis of federated learning systems under communication constraints.

B. Training Variance

The training variance in federated learning can come from different sources. A universal one is the variance of the stochastic gradient, which exists in all stochastic gradient decent (SGD)-based training algorithms. This variance is induced because in SGD-based training algorithms, instead of evaluating the true gradient $\nabla f(\mathbf{x})$ computed over the whole training dataset, which costs expensive computing resources, an estimation $\tilde{\nabla} f(\mathbf{x})$ is computed only over a mini-batch of the training dataset. Although $\tilde{\nabla} f(\mathbf{x})$ is usually an unbiased estimate of $\nabla f(\mathbf{x})$ [10], it has variance, given by $\mathbb{E}\|\tilde{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x})\|^2$, that depends on the size of the mini-batches. Compared to centralized SGD that only has the stochastic gradient variance, federated learning systems suffer from the variance induced by imperfect communication and data heterogeneity.

1) Data Heterogeneity: In practice, the local datasets $\mathcal{D}_i$, $i \in \{1, \cdots, N\}$, are drawn from possibly different and unknown distributions $\mathcal{P}_i$, $i \in \{1, \cdots, N\}$. Thus, the local objective functions, $f_i(\mathbf{x})$'s, are non-uniform (different) among the local devices and can deviate from the global objective $f(\mathbf{x})$. Thus, the presence of data heterogeneity renders the training and analysis of federated learning systems more challenging compared with traditional distributed learning systems.

2) Imperfect Communication: To see how imperfect communication results in additional training variance, we assume there is no stochastic gradient variance, i.e., $\tilde{\nabla} f(\mathbf{x}) = \nabla f(\mathbf{x})$, and examine the effects of the three communication-reduction techniques individually. First, when model compression is used, the gradient direction will be quantized as $Q(\nabla f(\mathbf{x}))$, which can deviate from $\nabla f(\mathbf{x})$. Although some stochastic compressors can provide an unbiased estimate of $\nabla f(\mathbf{x})$, the variance cannot be eliminated. Second, when the set of participating devices, $\mathcal{S}$, does not contain all the local devices, the aggregated gradient $\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x})$ may not align with $\nabla f(\mathbf{x})$. Third, under periodic aggregation, the devices can generate different local models that result in client drift [11].

3) Variance Reduction: Attempts at combating the training variance among local devices have led to the development of many variance-reduction techniques for federated learning. Some of them originate from DANE [12], which is a classical optimization method that introduces a sequence of local subproblems to reduce client drift. Fed-DANE [13] is adapted from DANE by allowing partial device participation. Network-DANE [14] is developed for decentralized federated learning. SCAFFOLD [11] can also be viewed as an improved version of DANE in federated settings by introducing some control variates.

C. Local Differential Privacy

Federated Learning achieves some levels of privacy by keeping the local datasets on user devices and only sharing the local updates with the server [15]. This, however, has been proven to be insufficient for maintaining data privacy as the parameters can reveal insights into the data that has been used for training. Consequently, FL by itself can only be incorporated with honest participating parties, and to extend it for secure and privacy-preserving settings, extra measures should be considered.

By design, federated learning is ignorant of how the local updates are being generated, making it vulnerable to different forms of privacy and robustness attacks from one or more malicious users [16]. Also, sharing the raw gradients can impose privacy risks for clients that can be exploited by a curious aggregation server, an adversary eavesdropping on the transmitted local updates, or a malicious participating client who might or might not be aware of the architecture of the model. A consequence of this is the membership inference attack which refers to an adversary with the intention of learning whether a certain record has been used for training the model (i.e., local private data of one of the clients) [17]. To this end, the adversary generates an attack model via its domain knowledge. Typically this is achieved by shadow models trained on noisy real or synthetic data or data acquired via model inversion attacks. As an example, one may assume a scenario of a model prescribing drugs for various patients. This form of attack allows the attacker to deduce if a patient participating in the study suffers from a disease (e.g., Alzheimer's), that is, sensitive information that could later be used against the patient. This phenomenon can be regarded as an implication of the model overfitting the training data which suggests using different regularization techniques (e.g., dropout and $l2$-norm, etc.) to overcome the problem. However, the high learning capacity of machine learning algorithms renders such attempts in preventing memorization of training data insufficient [18].

With the lack of a rigorous privacy guarantee for FL, there have been attempts to bring the de facto framework of privacy-preserving analysis, Differential Privacy (DP) [19], and its local counterpart, Local DP [20], to Federated Learning. Originally motivated by the inadequacy of anonymization techniques for enhancing privacy, DP has emerged as the gold standard of privacy protection. Privatizing machine learning by adding noise has introduced different methods based on the stage noise addition takes place. Specifically, for an Empirical Risk Minimization problem, three main approaches have been introduced for differentially private optimization: 1) *Objective Perturbation* where a randomized regularization term is added to the loss function, 2) *Output Perturbation* where noise is added to the parameters of a non-private model after training, and 3) the currently prevailing *Gradient Perturbation*, a more practical method that adds noise to the released gradient at each step, drawing more attention as it has been shown to be effective for nonconvex problems (as opposed to the last two methods) [21], [22].

To achieve data privacy in FL, instead of submitting raw local updates, the local parameters can be first perturbed using a randomization algorithm and then be released to the parameter server (local model), or alternatively, in the centralized fashion, the trusted parameter server can add noise to the aggregated updates (curator model). This addition of noise ensures that the local updates remain private and do not leak unnecessary information. Notice that there is a clear trade-off between the utility (accuracy of the model that is being trained) and the preserved privacy achieved by the noise. The perturbation is to impede attempts to infer the true values of a client with strong confidence but still allow accurate inferences for the population.

In the DP setting, it is assumed that the party responsible for aggregating the results, the parameter server, is trusted. Therefore the privacy could be maintained by adding noise to the aggregated results. However, for the surging edge computing and IoT applications, the parameter server ideally should not be trusted. This necessity naturally drives the research toward the local mode of DP, where each client would perturb its data to ascertain the data is kept private. This is to ensure that the clients' privacy is maintained even from the aggregator or an attacker that gets access to the data of the client on the server. However, the downside of the local mode is that the accumulated noise would incur more accuracy loss compared to the centralized context, mandating the need for more train data and longer training.

The privacy parameters (ε, δ) quantify DP where for smaller values thereof we get more privacy. Formally, a randomized algorithm $\mathcal{A}$ is (ε, δ) -differentially private if for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{A})$, and for all adjacent datasets D and D', then we have:

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\varepsilon} \Pr\left[\mathcal{A}\left(D'\right) \in \mathcal{S}\right] + \delta \tag{1}$$

where D and D' are two datasets that only differ in a single entry. In this context, the randomized algorithm $\mathcal{A}$ provides privacy by making the two datasets difficult to distinguish. However, in the absence of trust with the data collector, local DP is deemed more suitable. Formally, a randomized mechanism $\mathcal{M}$ is (ε, δ)-LDP, for any pair of inputs x and x' in $\mathcal{X}$, and any measurable subset $\mathcal{O} \subseteq \mathrm{Range}(\mathcal{M})$, then we have:

$$\Pr[\mathcal{M}(x) \in \mathcal{O}] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{M}\left(x'\right) \in \mathcal{O}\right] + \delta \tag{2}$$

Again, the privacy guarantee of $\mathcal{M}$ is determined by ε, but with a low probability of δ this might not hold.

D. Contributions

Although the effects of model compression, partial device participation, periodic aggregation, and data heterogeneity have been studied, a few works in the literature provide a comprehensive analysis by jointly considering all of them. This paper, to the best of our knowledge, is the first one presenting the convergence analysis of stochastic gradient descent in privacy-preserving federated settings by considering all the three communication-reduction techniques under the presence of data heterogeneity for strongly convex loss functions. Based on the results, we can have a clear understanding of how these components affect the convergence of federated learning systems and interplay with each other. Furthermore, we investigate differential privacy in the context of federated learning by introducing privacy-augmented FedPaq and discussing the impact of parameters describing the gradient perturbation on the performance of the model. Moreover, a privacy amplification method, based on subsampling of local datasets, is employed to enhance the convergence rate of the privatized model. In addition, our analysis provides practical insights into the design of a privatized FL model and communication strategies to accelerate the training process of such systems under a limited communication budget. To be specific, since all the parameters involved in the privacy measure and the three communication-reduction techniques are jointly considered, we are able to quantitatively analyze the trade-offs between different options. As a result, important design intuitions for real-world differentially private federated learning systems that are limited by the communication capacity constraints in wireless networks are provided.

E. Related Works

There have been several works on the convergence analysis of federated learning systems with different communication-reduction approaches.

The convergence rate of the FedAvg algorithm, which was first introduced in [3], has been studied in a non-i.i.d. data setting with partial device participation and periodic aggregation in [23]. However, the authors did not consider quantization. Similar set of results were presented in [24]. In [25], the FedPaq framework was introduced and analyzed under all three communication-reduction approaches. However, the authors assumed i.i.d. (homogeneous) data. To combat the effect of data heterogeneity, [26], [27] introduced an adaptation of FedAvg, named FedProx. In FedProx, a proximal term is introduced to each local objective function. However, these two works did not consider quantization. Finally, and perhaps the most recent work, [11] introduced the SCAFFOLD framework that incorporates a variance-reduction mechanism to combat the effect of non-i.i.d. data. More specifically, in

this framework, some control variates are introduced to reduce the drifts among different devices. While in [11], the authors show SCAFFOLD achieves better performance compared with FedAvg, the performance improvement is not significant under moderate heterogeneity (e.g., 10% similarity). On the other hand, since in SCAFFOLD all active devices need to send an additional control variate update, which has the same dimension as the model parameter, to the parameter server, the communication cost is doubled in each round. In practice, FedPaq is still a good candidate federated learning algorithm. However, in the original paper [25], the convergence analysis relies on the i.i.d. data assumption, which is unrealistic. To investigate the trade-offs between communication costs and training variance, it is important to obtain the convergence results for FedPaq under data heterogeneity, which is one of the main contributions of this paper.

In the literature, various randomized mechanisms and variations of DP and LDP, referred to as protocols, have been investigated to bring privacy guarantees to FL. In [28], the notion of Condensed LDP ($\alpha-$CLDP) is introduced which is later used in [29] to propose LDP-FL. Developing quantization schemes that are more compatible with continuous Gaussian and Laplacian noise is a recent theme of the research. Adding continuous noise after quantization would turn the value into a continuous number and the benefits of compression are lost. For LDP, biased compressors can not be used as they break the independence between rounds. This has led to works on discrete noise addition and compressors that take privacy perturbation into account. cpSGD overcomes the prohibitive problem of discrete values by adding noise drawn from a Binomial distribution and showing that for small ε it mimics the Gaussian mechanism [30]. In [31], a discrete Gaussian noise is introduced which has been used in [32].

II. Resource-Constrained Privacy Augmented Federated Learning

We start by introducing a general framework of federated learning systems, which can be seen in Fig. 1. The system consists of N local devices located in a wireless network, all connected to a parameter server that coordinates all devices to train a shared model. In a given communication round $k \in [K]$, the parameter server first distributes the aggregated information $(\mathbf{x}_k, \mathbf{c}_k)$, which is obtained through last communication round, to all active devices, where $\mathbf{x}_k$ is the global model and $\mathbf{c}_k$ is the global control variate[1]. On the other hand, for a device i that belongs to the set of participating (active) devices, $\mathcal{S}_k$, it will calculate the local model update $\Delta \mathbf{x}_k^{(i)} := \mathcal{V}_i(\mathbf{x}_k, \mathbf{c}_k; \mathcal{D}_i)$ and the update of the control variate $\Delta \mathbf{c}_k^{(i)}$, for some local functions $\mathcal{V}_i, i \in \mathcal{S}_k$. The complete training process is summarized in Algorithm 1.

A. Federated Learning Algorithms

Different designs of $\mathcal{V}_i$ and $\mathcal{S}_k$ lead to different algorithms. Here, we introduce some exemplary algorithms:

[1]We use $[K]$ to represent the set $\{1, \cdots, K\}$.

Algorithm 1 General Federated Learning

Input: global learning rate $\eta_{k,g}$ for $k \in [K]$
Initialize: model parameters $\mathbf{x}_0 \in \mathbb{R}^d$
for each round $k = 0, \cdots, K-1$ **do**
  on each device $i \in \mathcal{S}_k$:
    calculate $\Delta \mathbf{x}_k^{(i)} = \mathcal{V}_i(\mathbf{x}_k, \mathbf{c}_k; \mathcal{D}_i)$
    (optional) calculate $\Delta \mathbf{c}_k^{(i)}$
    send $\Delta \mathbf{x}_k^{(i)}$ and $\Delta \mathbf{c}_k^{(i)}$ to the parameter server
  on the parameter server:
    collect the local updates from devices in $\mathcal{S}_k$
    calculate $\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\eta_{k,g}}{M} \sum_{i \in \mathcal{S}_k} \Delta \mathbf{x}_k^{(i)}$
    (optional) calculate $\mathbf{c}_{k+1}$
    broadcast $\mathbf{x}_{k+1}$ and $\mathbf{c}_{k+1}$ to all devices.
end for

1) Distributed SGD: In distributed SGD [33], all devices participate in each communication round, i.e., $\mathcal{S}_k = [N]$ for all $k \in [K]$. Additionally, the local model update is calculated by one local SGD step, such that

$$\Delta \mathbf{x}_k^{(i)} = -\eta_{k,l} \tilde{\nabla} f_i(\mathbf{x}_k). \tag{3}$$

where $\eta_{k,l}$ is the local learning rate used at communication round k. Distributed SGD does not include a control variate. However, since it requires full participation and communication at each training round, it can result in overwhelming communication costs.

2) FedAvg and FedPaq: FedAvg [23] is proposed to save communication costs by using partial participation and periodic aggregation. Let E be the number of local iterations, then a device $i \in \mathcal{S}_k$ generates a sequence of local models $\left\{\mathbf{y}_{k,0}^{(i)}, \mathbf{y}_{k,1}^{(i)}, \cdots, \mathbf{y}_{k,E}^{(i)}\right\}$ with

$$\mathbf{y}_{k,0}^{(i)} = \mathbf{x}_k \quad \text{and} \quad \mathbf{y}_{k,t+1}^{(i)} = \mathbf{y}_{k,t}^{(i)} - \eta_k^l \tilde{\nabla} f_i\left(\mathbf{y}_{k,t}^{(i)}\right), \tag{4}$$

for $t \in [E]$. Then, the local model update is given by

$$\Delta \mathbf{x}_k^{(i)} = \mathbf{y}_{k,E}^{(i)} - \mathbf{x}_k. \tag{5}$$

Similarly, FedPaq [25] also generates the sequence of local models in (4), but clients would send a quantized version of local updates to the parameter server, i.e.,

$$\Delta \mathbf{x}_k^{(i)} = Q\left(\mathbf{y}_{k,E}^{(i)} - \mathbf{x}_k\right) \tag{6}$$

to further reduce communication overhead. Notice that FedPaq is a generalization of both distributed SGD (with $E = 1$, $\mathcal{S}_k = [N]$, and $Q(\mathbf{x}) = \mathbf{x}$) and FedAvg (with $Q(\mathbf{x}) = \mathbf{x}$).

3) SCAFFOLD: Instead of updating the local models only by the local stochastic gradients $\nabla f_i(\mathbf{y})$, SCAFFOLD [11] uses some control variates to reduce the drifts that occur among different devices. To be specific, one local step in SCAFFOLD is given by

$$\mathbf{y}_{k,t+1}^{(i)} = \mathbf{y}_{k,t}^{(i)} - \eta_k^l \left(\tilde{\nabla} f_i\left(\mathbf{y}_{k,t}^{(i)}\right) - \mathbf{c}_k^{(i)} + \mathbf{c}_k\right), \tag{7}$$

while the local control variate $\mathbf{c}_k^{(i)}$ and the global control variate $\mathbf{c}_k$ are updated by

$$\mathbf{c}_{k+1}^{(i)} = \mathbf{c}_k^{(i)} - \mathbf{c}_k - \frac{1}{E\eta_k^l} \Delta \mathbf{x}_k^{(i)}, \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

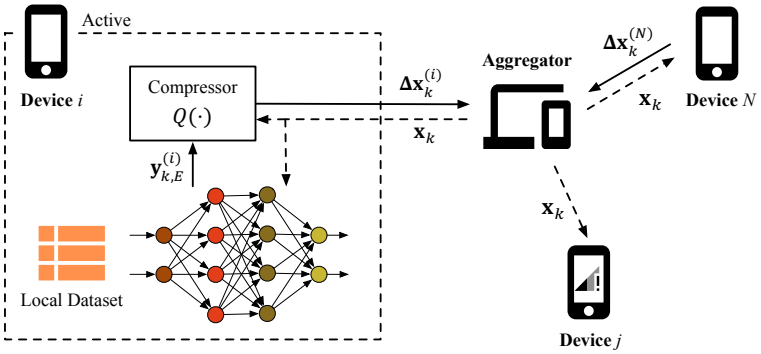


Fig. 1: Federated learning system.

$$\mathbf{c}_{k+1} = \mathbf{c}_k + \frac{1}{N}\sum_{i\in\mathcal{S}_k}\left(\mathbf{c}_{k+1}^{(i)} - \mathbf{c}_k^{(i)}\right). \tag{9}$$

The idea behind SCAFFOLD is to mimic the ideal update under centralized SGD, i.e.,

$$\tilde{\nabla} f_i\left(\mathbf{y}_{k,t}^{(i)}\right) - \mathbf{c}_k^{(i)} + \mathbf{c}_k \approx \frac{1}{N}\sum_{j\in[N]}\tilde{\nabla} f_j\left(\mathbf{y}_{k,t}^{(j)}\right). \tag{10}$$

B. Communication Constraints

To analyze the trade-offs between communication costs and training variance, we focus our analysis on FedPaq, which jointly considers model compression, partial device participation, and periodic aggregation under data heterogeneity. To simplify the analysis while preserving the essence of the communication constraint, we assume the capacity of the underlying M-user MAC channel for the federated learning system is bounded by $\mathcal{C}$ bits/second [34] and the total duration of the training process is $\mathcal{T}$ seconds. During the training process, each local device can conduct a total of T training iterations while the total of $B = \mathcal{C}\mathcal{T}$ bits can be shared among the active devices participating in the model aggregation process. Accordingly, we can link the communication constraints of the federated learning system of interests in the following:

$$B = \mathcal{C}\mathcal{T} = KM\beta = \left\lfloor\frac{T}{E}\right\rfloor M\beta \approx \frac{TM\beta}{E}, \tag{11}$$

where $K = \left\lfloor\frac{T}{E}\right\rfloor$ is the total number of training rounds and β is the number of bits required to transmit the model update. In this way, we can provide a unified framework to conduct performance analysis of federated learning under communication constraints.

C. Privacy Measure

Algorithm 2 outlines the introduced privacy-aware version of FedPaq. After E local iterations, the clients generate the local updates which are to be sent to the parameter server. Prior to uploading the local model updates, they get perturbed by a random noise drawn from a Gaussian distribution to preserve the privacy level of desire, i.e., a well-known procedure referred to as the Gaussian mechanism for achieving (ε, δ)-DP privacy guarantee. To locally differentially privatize a function $\mathrm{f}(X)$ subject to (ε, δ) we use

Algorithm 2 Privacy Augmented FedPaq.

Input: η_k for $k \in [K]$
Initialize: model parameters $\mathbf{x}_0 \in \mathbb{R}^d$
for each round $k = [K]$ **do**
  on each device $i \in \mathcal{S}_k$:
    Initialize the local model $\mathbf{x}_{k,0}^{(i)} = \mathbf{x}_k$
    Select a random subset $\mathcal{D}_s$ of size Eb from $\mathcal{D}_i$
    for each iteration $t = 0, \cdots, E-1$ **do**
      calculate $\mathbf{x}_{k,t+1}^{(i)} = \mathbf{x}_{k,t}^{(i)} - \eta_k \tilde{\nabla} f_i\left(\mathbf{x}_{k,t}^{(i)}\right)$
    end for
    $\mathbf{z}_k^{(i)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{i,k}^2 \mathbf{1}_d\right)$
    send $\Delta\mathbf{x}_k^{(i)} = Q\left(\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k + \mathbf{z}_k^{(i)}\right)$ to the server
  on the parameter server:
    calculate $\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{1}{M}\sum_{i\in\mathcal{S}_k}\Delta\mathbf{x}_k^{(i)}$
    broadcast the global model $\mathbf{x}_{k+1}$ to all devices
end for

$$M(X, \mathrm{f}, \sigma) \triangleq \mathrm{f}(X) + \mathcal{N}\left(0, \sigma^2\mathbf{I}\right), \tag{12}$$

with

$$\epsilon = \frac{\Delta_f}{\sigma}\sqrt{2\ln\frac{1.25}{\delta}}, \tag{13}$$

for any $\delta \in (0, 1]$ where Δ_f bounds the L_2 sensitivity of $\mathrm{f}(X)$, that is

$$\|f(x) - f(x')\|_2 \leq \Delta_f, \forall x, x' \in X \tag{14}$$

To reduce the amount of noise required to be added to the local updates to achieve a certain privacy guarantee, one may use privacy amplification techniques. Here we employ subsampling into mini-batches to achieve this reduction. An algorithm f would reach a better privacy guarantee when applied on a random subsample of the dataset instead of the full sample. This is intuitively due to the fact that data not included in the subsample would enjoy full privacy, hence the privacy being amplified. Applying the (ϵ, δ)-DP randomized mechanism $\mathcal{M}$ on the subsampled data achieves $(\epsilon', h(\delta))$-DP mechanism $\mathcal{M}^{\mathcal{S}}$ for $0 \leq \epsilon' \leq \epsilon$ and a function h to

be determined based on the sampling procedure[2] [35]. For subsampling E mini-batches of size b without replacement we achieve $\left(\log\left(1+\left(1-(1-b/n_k)^E\right)(e^\epsilon-1)\right), \gamma\delta\right)$-DP for $\Delta\mathbf{x}_k^{(i)}$ where $\gamma := Eb/n_k$.

Notice that a (ϵ, δ)-DP mechanism $\mathcal{M}$ is also (ϵ', δ') for any $\epsilon' \geq \epsilon$ and any $\delta' \geq \delta$. Then,

$$\begin{aligned}&\log\left(1+\left(1-(1-b/n_k)^E\right)(e^\epsilon-1)\right)\\ &\leq \log\left(1+\gamma\left(e^\epsilon-1\right)\right) \leq \gamma\left(e^\epsilon-1\right) \leq 2\gamma\epsilon,\end{aligned} \tag{15}$$

hence the introduced method satisfying at least (ϵ, δ)-DP for $\mathbf{x}_{k,E}^{(i)}$ via adding reduced Gaussian noise of

$$\sigma_{i,k}^2 = \frac{2\left(\Delta_f\right)^2 \ln(1.25/(\delta/\gamma))}{(\epsilon/2\gamma)^2}. \tag{16}$$

Following this procedure, having fixed the privacy parameters ε and δ, the algorithm achieves higher utility (in terms of model accuracy) and converges faster. Needless to say, clipping the local gradients to a small value to maintain lower global sensitivity, and subsampling, although result in noise of less magnitude, also may inversely impact the convergence in the non-secure setting. Therefore, careful convergence analysis is deemed necessary.

III. Convergence of Privatized FedPaq Under Heterogeneity

In this section, we present the convergence analysis for the Privatized FedPaq algorithm under non-i.i.d. data setting (Algorithm 2) with the following assumptions:

Assumption 1: The loss function ℓ is L-smooth, such that $\|\nabla\ell(\mathbf{x}) - \nabla\ell(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Consequently, f_i and f are also L-smooth.

Assumption 2: The loss function ℓ is μ-strongly convex, such that $\|\nabla\ell(\mathbf{x}) - \nabla\ell(\mathbf{y})\| \geq \mu\|\mathbf{x}-\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Consequently, f_i and f are also μ-strongly convex.

Assumption 3: The stochastic gradient is unbiased and variance-bounded, such that for any $\mathbf{x} \in \mathbb{R}^d$ and the mini-batch samples ξ, we have $\mathbb{E}\left[\nabla f_i\left(\mathbf{x};\xi\right)\right] = \nabla f_i\left(\mathbf{x}\right)$ and $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x};\xi\right) - \nabla f_i(\mathbf{x})\right\|^2 \leq \sigma^2/b$ for all $i = 1, 2, \cdots, N$. Here, b is the mini-batch size.

Assumption 4: $Q(\cdot)$ is an unbiased and q-lossy random compressor, such that for any $\mathbf{x} \in \mathbb{R}^d$ $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}$ and $\mathbb{E}\|Q(\mathbf{x}) - \mathbf{x}\|^2 \leq q\|\mathbf{x}\|^2$. Note that q is a real value in $[0, 1]$, which is determined by the resolution of the compressor, β. When $q = 0$, we have $Q(\mathbf{x}) = \mathbf{x}$ and there is no quantization. For $q = 1$, no information is being sent during communications.

Remark 1: While Assumptions 1, 2, and 3 are widely used in literature, the unbiasedness of the compressed model stated Assumption 4 does not always hold in practice. For example, the SignSGD [36] uses biased quantizer which makes the analysis more complicated. For the tractability of our analysis, we use an unbiased compressor throughout this paper. Examples for unbiased quantizers include quantized-SGD and TernGrad [37], [38], which preserve the true values in expectation. Furthermore, for quantized-SGD, [37] shows that $q = \min\{n/s^2, \sqrt{n}/s\}$, where n is the block-size and s is the quantization level. For simplicity, we assume $n = d$.

Remark 2: While the assumption of strong convexity is restrictive, it facilitates our analysis since the convergence rate can be directly measured by $\|\mathbf{x}_k - \mathbf{x}^*\|^2$. Additionally, $\mu > 0$ allows us to use the learning rate $\eta_k = \frac{4\mu^{-1}}{kE+4E}$ in Theorem 1. Relaxing this assumption is part of our future work.

Remark 3: Notice that Assumption 3 is general for all choices of batch size when calculating the stochastic gradients. A larger batch size will lead to a smaller σ.

Remark 4: Notice that Assumptions 1, 2, and 3 are also used in [23], which proves the convergence rate for FedAvg under non-iid data. However, our paper differs from [23] in that we consider privacy concerns and compression of model updates, making our system more realistic. However, this also makes the underlying analysis more challenging where the original results cannot be directly applied. Moreover, [23] has an additional assumption that the expected squared norm of stochastic gradients is uniformly bounded, i.e., $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,t}^{(i)}; \xi_{k,t}^{(i)}\right)\right\|^2 \leq G^2$, where $G > 0$, for all devices and all time steps.

Assumption 5: We assume that for all $\mathbf{x} \in \mathbb{R}^d$, there exist a $\lambda > 0$, such that

$$\frac{1}{N}\sum_{i\in[N]}\left\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq \lambda^2. \tag{17}$$

The value of λ quantifies the degree of data heterogeneity among different local devices. When $\lambda = 0$, we have $\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$ for all $i = 1, 2, \cdots, N$, and there is no heterogeneity. It can be seen from (17) that larger λ implicates higher degree of heterogeneity.

Remark 5: Different papers usually have different assumptions on data heterogeneity. For example, [23] defines $\Gamma = f^* - \sum_{i=1}^N w_i f_i^*$ and [24] defines $\varsigma^2 = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x^*)$ for quantifying the degree of non-i.i.d. Although our measure of heterogeneity is different from these two papers, we note that both Γ and ς can still be expressed using λ under strongly convex setting.

Assumption 6: The stochastic gradient is uniformly bounded, such that for any $\mathbf{x} \in \mathbb{R}^d$ and the mini-batch samples ξ, we have $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x};\xi\right)\right\|^2 \leq G^2$ for all $i = 1, 2, \cdots, N$.

For the simplicity of our analysis, we additionally assume the device weights, w_i's, to be uniform among all devices. However, as suggested in [23], this does not result in the loss of generalization of our results. Indeed, by replacing the local objective as $\tilde{f}_i(\mathbf{x}) = w_i N f_i(\mathbf{x})$ and transforming the value of L, μ, σ, and λ accordingly, the global objective becomes the simple average of transformed local objectives, i.e., $f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \tilde{f}_i(\mathbf{x})$. Moreover, we assume the same gradient perturbation is being performed by all clients.

[2]Poisson subsampling, sampling without replacement and sampling with replacement are possible options that have been studied in the literature.

Using the assumptions and lemmas above, the convergence rate of Algorithm 2 for the secure federated learning setting can be characterized in the following theorem:

Theorem 1: Training the secure federated learning system using Algorithm 2, under Assumptions 1, 2, 3, 4, 5 and 6 for T iterations and setting the learning rate in the k-th round as $\eta_k = \frac{4\mu^{-1}}{kE+4E}$, the aggregated global model $\mathbf{x}_K$, where $K = \left\lfloor \frac{T}{E} \right\rfloor$, satisfies

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_K - \mathbf{x}^*\|^2 \leq & \frac{16E^2}{T^2}\mathbb{E}\|\mathbf{x}_{k_0} - \mathbf{x}^*\|^2 \\ & + \frac{16}{\mu^2}\left(\frac{2qG^2}{M} + \frac{qG^2}{N}\right)\frac{E}{T} \\ & + \frac{16}{\mu^2}\left(\frac{4e\sigma^2}{bM} + \frac{3L\lambda^2}{\mu} + \frac{\sigma^2}{bN}\right)\frac{1}{T} \\ & + \frac{128e\lambda^2}{\mu^2 M}\cdot\frac{E-1}{T} + \frac{128G^2}{\mu^2}\frac{(E-1)^2}{T} \\ & + \frac{4096dG^2b^2(1+q)\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{M|\mathcal{D}_i|^2\epsilon^2}\frac{E^3}{T}\end{aligned} \tag{18}$$

Remark 6: For vanilla version of the algorithm, without addition of noise required for privacy, similar to [23], [25], and [26], our result suggests a convergence rate as $\mathcal{O}\left(\frac{E^2}{T}\right)$, which requires $E = o\left(\sqrt{T}\right)$. However, this term appears only when data heterogeneity exists, i.e., $\lambda > 0$. When there is no data heterogeneity among local devices, the training can get significantly accelerated.

Remark 7: Convergence gets slowed down when q, E and λ become larger, whereas larger M (more active devices) can facilitate the convergence. However, since $\mathcal{O}\left(\frac{1}{M}\right)$ does not appear in all terms, the training does not enjoy a linear speedup.

Remark 8: The convergence rate is also impacted by other factors such as ϵ, δ and G. If ϵ and δ are small, the model convergence will slow down, thus necessitating more global iterations. Meanwhile, if G increases, the right part of (18) will also increase accordingly, and thus delays the model convergence.

A. Impacts of Communication Constraints

As shown in (11), various communication overhead reduction strategies will be coupled together under the constraint of MAC capacity. Therefore, it is important for us to characterize the impacts on the various trade-offs among model compression, partial participation, and periodic aggregation based on the conducted convergence analysis to obtain new system design intuitions for federated learning systems under communication constraints and differential privacy.

1) Partial Participation v.s. Periodic Aggregation: Assuming the model compression strategy is fixed, we can write $M = \alpha E$, where $\alpha = B/T\beta$. Furthermore, we assume a large-scale federated learning system where $N \gg M$, such that the approximation of $\frac{(N-M)(N-1)}{MN^2} \to \frac{1}{M}$ holds. When the number of total iterations T is large enough, the dominating term in (18) becomes

$$\mathcal{O}\left(\frac{aE^2 + b\lambda^2 + c\lambda^2/E + d}{T}\right), \tag{19}$$

where a, b, c, and d are positive constants determined by q, L and μ, along the the global sensitivity and privacy parameters. As it is evident from (19), in presence of data heterogeneity ($\lambda > 0$), increasing E can improve convergence for larger λ.

2) Model Compression v.s. Others: Consider the model compressor used in quantized-SGD [37], such that for any $\mathbf{x} \in \mathbb{R}^d$, the i-th element is quantized as

$$Q_i(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(\mathbf{x}_i) \cdot \vartheta_i(\mathbf{x}, s), \tag{20}$$

where $\vartheta_i(\mathbf{x}, s)$ is a random variable taking on value $\frac{l+1}{s}$ with probability $\frac{|\mathbf{x}_i|}{\|\mathbf{x}\|}s - l$ and $\frac{l}{s}$ otherwise. Here, $s \in \mathbb{Z}^+$ is the quantization level and $l \in [0, s)$ is an integer such that $\frac{|\mathbf{x}_i|}{\|\mathbf{x}\|} \in \left[\frac{l}{s}, \frac{l+1}{s}\right)$. To transmit $Q(\mathbf{x})$, three parts of information need to be encoded, including $\|\mathbf{x}\|^2$, $\{\text{sign}(\mathbf{x}_i)\}_{i=1}^d$, and $\{\vartheta_i(\mathbf{x}, s)\}_{i=1}^d$. Assuming we are using the simplest one-hot encoder, jointly encoding $\{\text{sign}(\mathbf{x}_i)\}_{i=1}^d$ and $\{\vartheta_i(\mathbf{x}, s)\}_{i=1}^d$ will cost $d\log_2(2s+1)$ bits, while $\|\mathbf{x}\|$ costs 32 bits. Assume $d\log_2(2s+1) \gg 32$, we have $\beta \approx d\log_2(2s+1)$. By taking $q = \frac{\sqrt{d}}{s}$, the quantization loss q is approximately $\mathcal{O}\left(\frac{2\sqrt{d}}{2^{\beta/d}-1}\right)$. Clearly, q will decrease rapidly with β. Furthermore, as we can see in Theorem 1, there is no term with q and λ coexisting, indicating that the impact of q is invariant to non-i.i.d. data. On the other hand, $\frac{E}{M} \sim \mathcal{O}(\beta)$ and the impact of E is magnified under data heterogeneity. Thus, we can choose a relatively small β for a good trade-off between model compression and the other two communication-reduction approaches.

3) Impact of Privacy Measure: As mentioned before, the local updates are clipped such that the global sensitivity is bounded. Theorem 1 suggests that the convergence can be severely delayed for large G. This is due to the fact that in 16, the noise variance quadratically grows with the L_2 sensitivity.

Increasing the subsampling ratio γ, determined by Eb/n_k, amplifies the noise variance required for desired privacy guarantee. For large E, the last terms in (18),

$$\frac{4096dG^2b^2(1+q)\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{M|\mathcal{D}_i|^2\epsilon^2}\frac{E^3}{T}$$

dominates convergence. In this case, increasing b, which consequently increases γ, would deteriorate the performance, hence our decision to employ small subsampling ratios to achieve privacy amplification. On the other hand, for small E increasing γ may not be as detrimental, or can even be beneficial, as is the case with the non-privatized algorithm. Clearly, increasing ϵ and δ improve the convergence rate, but it is not desirable as we would be opting for a risky and insufficient privacy guarantee.

IV. Performance Evaluation

1) Model and Dataset: The theoretical results are evaluated via a logistic regression model based on the widely-known MNIST dataset. The MNIST dataset is equally distributed among $N = 100$ local devices. We control the degree of data heterogeneity by allowing a local device to have access only to training samples for a fraction of all the 10 digits. Consequently, we can generate ten datasets `HET_MNIST(n_digits)`, for `n_digits` $= 1, 2, \cdots, 10$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

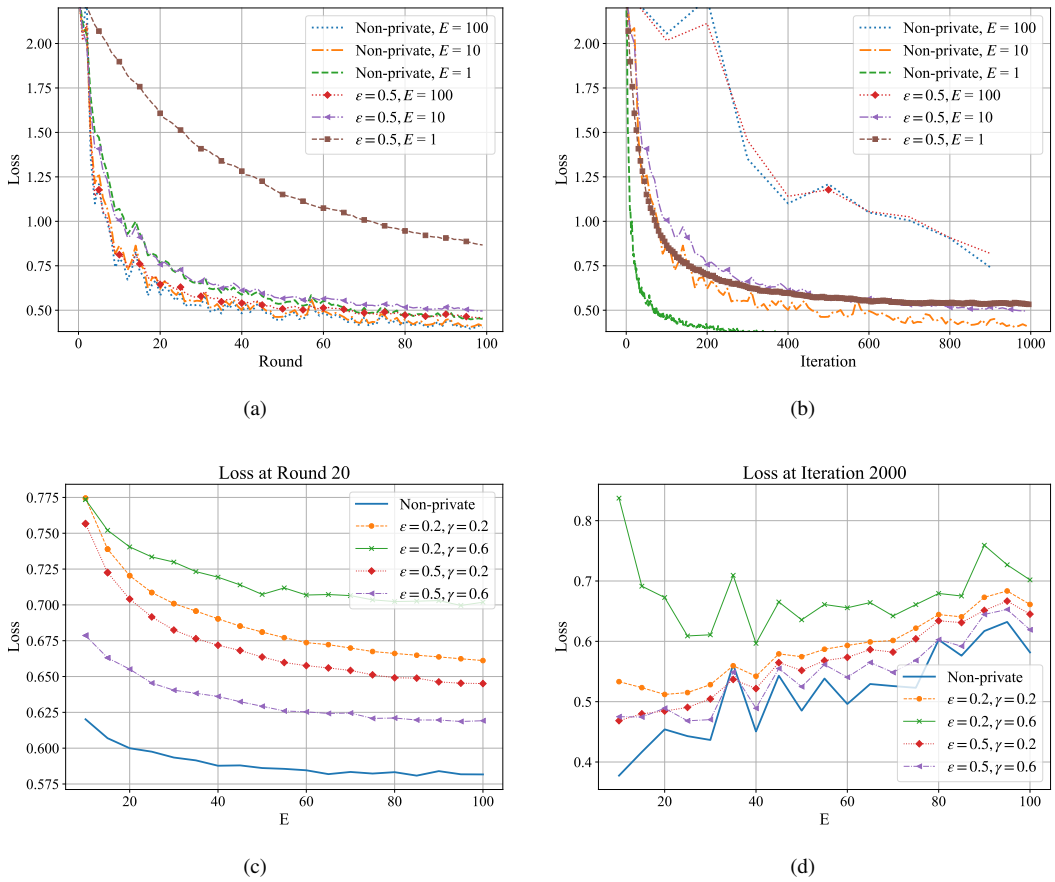


Fig. 2: The impact of E on `HET_MNIST`(2) dataset: (Top row): The training curves for different E with $M = 10$, $s = 10$, $\gamma = 0.2$ and $C = 1$. (a) Over rounds (b) Over iterations. (Bottom row) The training loss (a) after 20 communication rounds (b) after 2000 iterations.

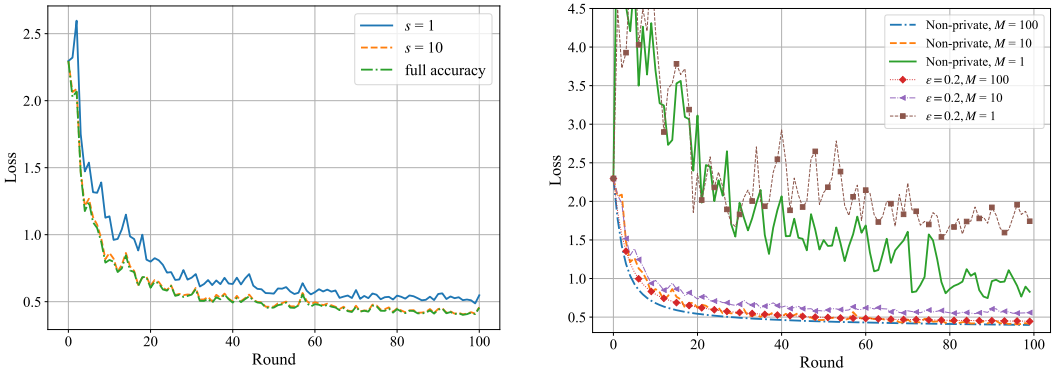
Clearly, choosing a smaller number of digits results in a higher degree of data heterogeneity, while `n_digits` $= 10$ indicates no heterogeneity among the local datasets.

2) Experiment Settings: The local models are aggregated for every E iterations, such that there are $K = \lfloor \frac{T}{E} \rfloor$ communication rounds. Unless otherwise stated, $E = 10$ local iterations are performed by each client, and the number of communication rounds K is fixed to 100, with the total number of iterations T accordingly calculated. The learning rate at the k-th round is set to be $\eta_k = \frac{\eta_0}{1+kE/100}$, where $\eta_0 = 0.1$. The non-privatized setting refers to the vanilla FedPaq algorithm, with neither subsampling, clipping, nor gradient perturbation. In the secure mode, a subset of the local datasets with cardinality γn_i is randomly selected without replacement by each participating device $i \in \mathcal{S}_k$ at the start of each round to be used for local training. During training, the per-sample gradients are clipped with respect to C, and proper noise is added prior to the transmission of local gradients. Upon aggregation, the scheduled devices will send the (noisy) model updates compressed by the quantizer in (20) with quantization level s. We evaluate the impact of communication-reduction techniques on both the non-privatized and secure aggregated models over training rounds. Furthermore, we will also investigate how each of the parameters involved in privatizing the Secure FedPaq, ceteris paribus, influences the performance of the introduced federated learning scheme.

3) Impact of Periodic Aggregation: Under the presence of data heterogeneity, the impact of E is illustrated in Fig. 2. We observe that the accuracy of the model increases for larger E per round, as illustrated in Fig. 2 (a). This refers to the realistic case of constraining the number of communication rounds to a fixed number. Alternatively, Fig. 2 (b) shows that fixing the number of global iterations T, larger E in fact delays the convergence of the model at each iteration. The empirical performance verifies our analytical result in Theorem 1, that for fixed T, increasing E would deteriorate the result. Intuitively, this means fewer opportunities for aggregating the local updates. However, proportionally increasing both E and T clearly helps both versions of FedPaq, although a diminishing gain is achieved. Fig. 2 (c) and (d) show the loss curves when the number of rounds or the number of global iterations is fixed, respectively. Fig. 2 confirms our interpretation of Theorem 1 that larger subsampling ratio γ would be more detrimental for larger values of E.

4) Impact of Partial Participation: As shown in Fig. 3(b), when M is set to a small value, for example $M = 1$, the training becomes extremely unstable during the first several

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal



(a) (b)

Fig. 3: Training curves for different s and M at each round on HET_MNIST(2) dataset: (a) Different s on the non-privatized model with $M = 10$ and $E = 10$. (b) Different M with $E = 10$, $s = 10$, $\gamma = 0.2$ and $C = 1$.

training rounds. Note that the impact is even more pronounced for the privatized algorithm, consistent with the last term of (18) where the impact of perturbation is inhibited for larger M.

5) Impact of Model Compression: In Fig. 3 (a), the loss curve for $s = 10$ is almost overlapped with the one without quantization. Even for $s = 1$, the performance degradation is small and the training is pretty stable. This motivates us to use lossy model compressors to save communication resources for smaller E and larger M, which have significant impacts on the convergence of the global model. The same impact can be observed for the privatized setting, which is omitted to avoid polluting the figure.

6) Impact of Privacy Budget ε: Increasing the privacy parameters ε and δ result in additive noise of lower magnitude, which brings more accurate estimation while putting the privacy of clients at risk. We have set $\delta = 10^{-4}$ across all the experiments. Fig. 4 (a) depicts the impact of ε on performance of the model. For $\varepsilon = 0.1$, although a high level of privacy guarantee is ensured, the utility significantly declines and convergence may be hindered.

7) Impact of Gradient Clipping: The global sensitivity G appears in the numerator of multiple terms in Eq. 18 suggesting its major impact on the convergence rate of the algorithm. Fig. 4 (b) shows that larger C, inducing more noise, delays convergence of the privatized algorithm. In fact, one might claim that a smaller C for clipping the gradients may be preferred to choosing larger ε for achieving faster convergence. However, based on Fig. 4 (d), this is not always true.

8) Impact of Subsampling: As a privacy amplification measure, we are using subsampling to reduce the amount of noise that needs to be added to maintain a certain level of privacy guarantee, thus increasing the utility. Fig. 2 (c) and (d) show how a larger subsampling ratio γ has a more pronounced deteriorating impact for greater values of E, verifying our analytical results. Moreover, Fig. 4 (c) depicts the training loss of a few privatized models for various values of γ, but otherwise identical. Fig. 4 (d) show how γ and C jointly affect the performance of the model. One can observe that a smaller clipping of $C = 0.5$ achieves the best or one of the worst privatized setting results based on which value of γ is chosen. Although small C and γ lead to relatively negligible added noise, however, clipping the gradients acquired on such small subsample would significantly deteriorate convergence, as would be the case with non-privatized SGD. A more moderate clipping, namely $C = 1$, would be more resilient in face potentially undesirable consequences of subsampling.

The theoretical and experimental results suggest some important design intuitions for federated learning systems: (1) The choice of model compression accuracy has little impact on the convergence regardless of the heterogeneity of the problem. This encourages the widespread use of low-accuracy quantizers to save the underlying communication overhead even in non-iid settings. (2) Although larger E naturally incurs more computational cost for the clients, we realized that in the non-iid setting, it could actually enhance the performance for a fixed number of communication rounds K. (3) A smaller M will reduce the convergence rate of the model. This behavior is even more noticeable in the privatized setting. (4) Subsampling becomes more important for larger E. (5) Generally, a lower sampling ratio can achieve better convergence, but it may make the optimization unstable if the global sensitivity is very low.

Finally, although our work on extending the results to the cases with more relaxed convexity assumptions is still in progress, our empirical simulations on applying the introduced scheme on privacy-aware federated learning of ANNs, CNNs, and Spiking Neural Networks indicate that having set the learning rate to a proper value, the same effects can be seen for hyperparameters involved in the FL design as presented in this paper. Furthermore, another interesting direction for future work is to explore similar designs that are consuming less energy, ranging from energy-efficient neuromorphic computing [39] to energy optimization when the FL design is incorporated for mobile-edge computing [40].

V. Conclusion

In this paper, we provided a comprehensive convergence analysis for a privacy augmented federated learning system by considering data heterogeneity as well as the communication

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

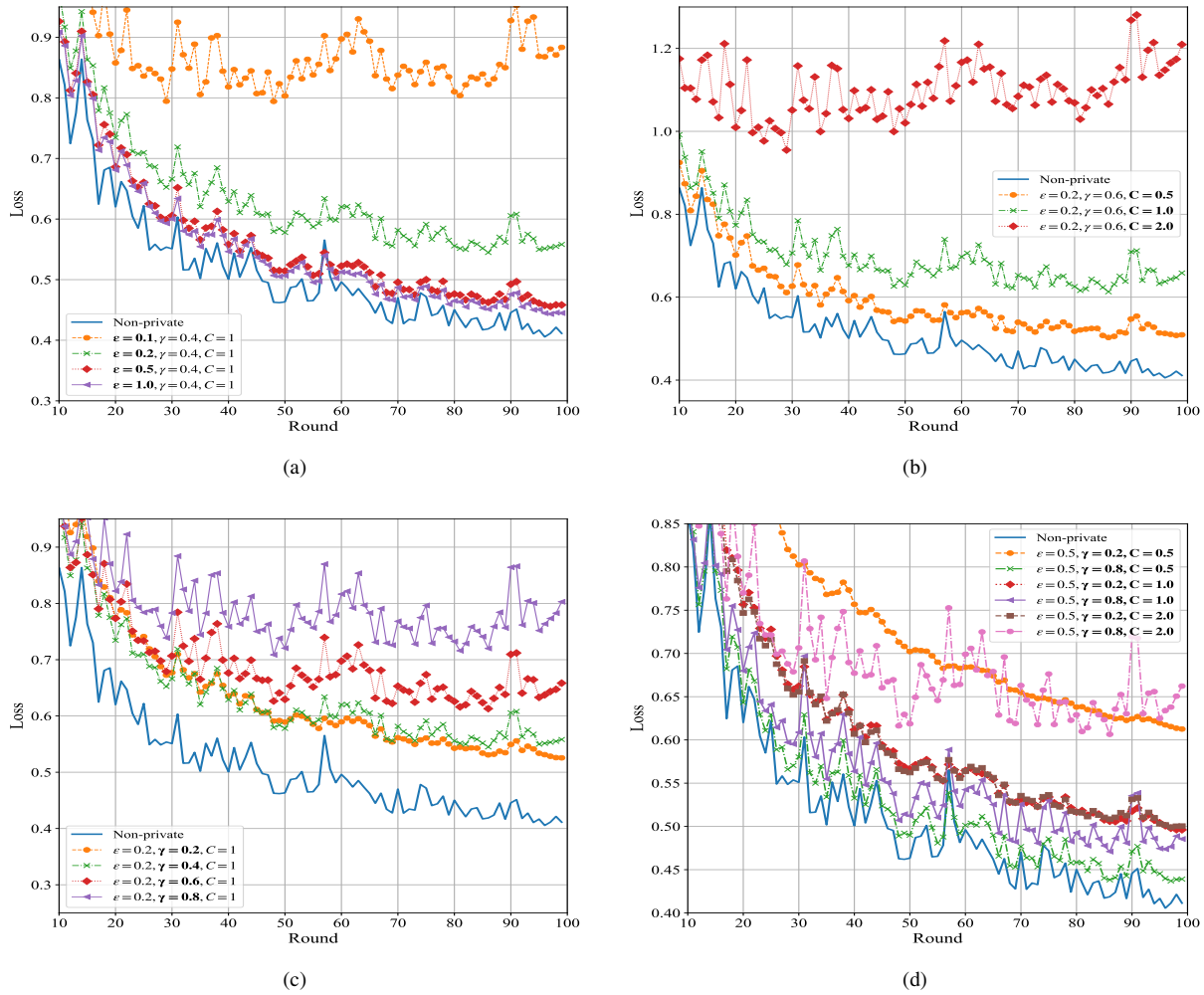


Fig. 4: The impact of different privacy-related parameters on convergence of the privatized FedPaq algorithm with $E = 10$, $M = 10$ and $s = 10$ (a) for different ε (b) Training curves of different C, (c) Training curves of different γ, (c) Training curves for both γ and C.

constraints. To be specific, three communication-reduction strategies, namely model compression, partial device participation, and periodic aggregation are jointly considered under the capacity limit of the underlying MAC for model aggregation. Due to the insufficiency of FL in maintaining the privacy of clients, a privacy measure based on differential privacy is considered, introducing the privacy-augmented FedPaq algorithm. The impacts of differential privacy, data heterogeneity and the communication-reduction strategies were evaluated both theoretically and numerically. Our analysis provides important design intuitions for real-world federated learning systems that are limited by the communication capacity constraints in wireless networks.

REFERENCES

[1] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, "Artificial intelligence-enabled cellular networks: A critical path to Beyond-5G and 6G," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 212–217, 2020.

[2] H. Song, J. Bai, Y. Yi, J. Wu, and L. Liu, "Artificial intelligence enabled internet of things: Network architecture and spectrum access," *IEEE Comput. Intell. Maga.*, vol. 15, no. 1, pp. 44–51, 2020.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[4] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4424–4434.

[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," *arXiv preprint arXiv:1807.00459*, 2018.

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

[8] A. D. Wyner, "Shannon-theoretic approach to a gaussian cellular multiple-access channel," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[10] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[11] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for on-device federated learning," *arXiv preprint arXiv:1910.06378*, 2019.

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

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smithy, "Feddane: A federated newton-type method," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1227–1231.

[14] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1662–1672.

[15] B. Shang, S. Liu, S. Lu, Y. Yi, W. Shi, and L. Liu, "A cross-layer optimization framework for distributed computing in iot networks," in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2020, pp. 440–444.

[16] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[18] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[19] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[20] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[21] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization." *Journal of Machine Learning Research*, vol. 12, no. 3, 2011.

[22] D. Yu, H. Zhang, and W. Chen, "Improve the gradient perturbation approach for differentially private optimization," 2018.

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

[24] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local gd on heterogeneous data," *arXiv preprint arXiv:1909.04715*, 2019.

[25] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," *arXiv preprint arXiv:1909.13014*, 2019.

[26] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[27] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "On the convergence of federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[28] M. E. Gursoy, A. Tamersoy, S. Truex, W. Wei, and L. Liu, "Secure and utility-aware data collection with condensed local differential privacy," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[29] L. Sun, J. Qian, X. Chen, and P. S. Yu, "LDP-FL: Practical private aggregation in federated learning with local differential privacy," *arXiv preprint arXiv:2007.15789*, 2020.

[30] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpsgd: Communication-efficient and differentially-private distributed sgd," in *Advances in Neural Information Processing Systems*, 2018, pp. 7564–7575.

[31] C. Canonne, G. Kamath, and T. Steinke, "The discrete gaussian for differential privacy," *arXiv preprint arXiv:2004.00010*, 2020.

[32] L. Wang, R. Jia, and D. Song, "D2p-fed: Differentially private federated learning with efficient communication," 2020.

[33] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous sgd," *arXiv preprint arXiv:1604.00981*, 2016.

[34] T. M. Cover and J. A. Thomas, *Elements of information theory (2nd Edition)*. John Wiley & Sons, 2012.

[35] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1226–1235.

[36] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signsgd: Compressed optimisation for non-convex problems," *arXiv preprint arXiv:1802.04434*, 2018.

[37] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[38] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in neural information processing systems*, 2017, pp. 1509–1519.

[39] R. Shafin, L. Liu, J. Ashdown, J. Matyjas, M. Medley, B. Wysocki, and Y. Yi, "Realizing green symbol detection via reservoir computing: An energy-efficiency perspective," in *2018 IEEE Intl Conf Communications (ICC)*, 2018, pp. 1–6.

[40] B. Shang and L. Liu, "Mobile-edge computing in the sky: Energy optimization for air–ground integrated networks," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7443–7456, 2020.

VI. Proofs

To facilitate the analysis, we introduce two quantities

$$\begin{aligned}\hat{\mathbf{x}}_{k+1} &= \mathbf{x}_k + \frac{1}{N}\sum_{i=1}^{N} Q\left(\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k + \mathbf{w}_k^{(i)}\right),\\ \overline{\mathbf{x}}_{k,E} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{k,E}^{(i)},\end{aligned} \tag{21}$$

where $\mathbf{w}_k^{(i)} \sim \mathcal{N}\left(0, \varsigma_{i,k}^2\mathbf{I}\right)$.

Additionally, we note as $\mathbf{x}^*$ and $\mathbf{x}_i^*$ the optimal global model and optimal local model for the i-th device, respectively. As shown in [25], for a fixed global iteration k, the squared difference between $\mathbf{x}_{k+1}$ and $\mathbf{x}^*$ can be decomposed into three terms

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = &\underbrace{\mathbb{E}\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\|^2}_{A}\\ &+ \underbrace{\mathbb{E}\|\hat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,E}\|^2}_{B}\\ &+ \underbrace{\mathbb{E}\|\overline{\mathbf{x}}_{k,E} - \mathbf{x}^*\|^2}_{C},\end{aligned} \tag{22}$$

which are bounded as shown in the following lemmas:

Lemma 1: Under Assumptions 1, 2, 3, 4, and 5, term A in (22) can be bounded as

$$\begin{aligned}\mathbb{E}\left\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\right\|^2 \leq &\frac{2qG^2(N-M)}{MN}E^2\eta_k^2\\ &+ \frac{4e\sigma^2(N-M)}{bMN}E\eta_k^2\\ &+ \frac{8e\lambda^2(N-M)}{MN}E(E-1)\eta_k^2\\ &+ \frac{256dG^2b^2(1+q)(N-M)\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{M(N-1)|\mathcal{D}_i|^2\epsilon^2}E^4\eta_k^2.\end{aligned} \tag{23}$$

Lemma 2: Under Assumptions 1, 2, 3, 4, and 5, term B in (22) can be bounded as

$$\mathbb{E}\|\hat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,E}\|^2 \leq \frac{qG^2}{N}E^2\eta_k^2. \tag{24}$$

Lemma 3: Under Assumptions 1, 2, 3, 4, and 5, term C in (22) can be bounded as

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,E} - \mathbf{x}^*\right\|^2 \leq &(1-\mu\eta_k)^E\mathbb{E}\left\|\mathbf{x}_k - \mathbf{x}^*\right\|^2\\ &+ \left(\frac{3L\lambda^2}{\mu} + \frac{\sigma^2}{bN}\right)E\eta_k^2\\ &+ 8G^2E(E-1)^2\eta_k^2.\end{aligned} \tag{25}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

A. Proof of Theorem 1

By substituting (23), (24), and (25) into the (22), we obtain

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 \leq&(1-\mu\eta_k)^E\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2\\&+D_1E^2\eta_k^2+D_2E\eta_k^2\\&+D_3E(E-1)\eta_k^2\\&+D_4E(E-1)^2\eta_k^2\\&+D_5E^4\eta_k^2,\end{aligned}\tag{26}$$

where

$$\begin{aligned}D_1=&\frac{2qG^2(N-M)}{MN}+\frac{qG^2}{N},\\D_2=&\frac{4e\sigma^2(N-M)}{bMN}+\frac{3L\lambda^2}{\mu}+\frac{\sigma^2}{bN},\\D_3=&\frac{8e(N-M)\lambda^2}{MN},\\D_4=&8G^2,\\D_5=&\frac{256dG^2b^2(1+q)(N-M)\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{M(N-1)|\mathcal{D}_i|^2\epsilon^2}.\end{aligned}\tag{27}$$

Let $\eta_k \leq \frac{1}{\mu}$. Thus, the term $(1-\mu\eta_k)^E$ can be bounded as

$$\begin{aligned}(1-\mu\eta_k)^E=&\left(1-\frac{\mu E\eta_k}{E}\right)^E\\\overset{(a)}{\leq}&e^{-\mu E\eta_k}\\\overset{(b)}{\leq}&1-\mu E\eta_k+\frac{1}{2}\mu^2E^2\eta_k^2,\end{aligned}\tag{28}$$

where (a) holds because $(1+\frac{x}{n})^n \leq e^x$ for $|x|<n$, and (b) holds because $e^x \leq 1+x+\frac{x^2}{2}$ for $x \leq 0$. Then, for $(1-\mu\eta_k)^E$, we have

$$(1-\mu\eta_k)^E \leq 1-\mu E\eta_k+\frac{1}{2}\mu^2E^2\eta_k^2.\tag{29}$$

By restricting

$$\eta_k \leq \frac{1}{\mu E},\tag{30}$$

it implies

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_{k+1}-\mathbf{x}^*\| \leq&\left(1-\frac{1}{2}\mu E\eta_k\right)\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2\\&+D_1E^2\eta_k^2+D_2E\eta_k^2\\&+D_3E(E-1)\eta_k^2\\&+D_4E(E-1)^2\eta_k^2\\&+D_5E^4\eta_k^2.\end{aligned}\tag{31}$$

By setting the learning rate to $\eta_k=\frac{4\mu^{-1}}{kE+\gamma_0}$ (a diminishing learning rate) and defining $\gamma:=\frac{\gamma_0}{E}$ and $\delta_k:=\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2$, we can write the above expression as

$$\delta_{k+1} \leq \left(1-\frac{2}{k+\gamma}\right)\delta_k+\frac{\alpha}{(k+\gamma)^2},\tag{32}$$

where

$$\alpha=\frac{16}{\mu^2}\left(D_1+\frac{1}{E}D_2+\frac{E-1}{E}D_3+\frac{(E-1)^2}{E}D_4+E^2D_5\right)\tag{33}$$

Now set $\gamma_0=\max\left\{4E,\frac{4L^2}{\mu^2},\frac{4\sqrt{2}L(E-1)}{\mu}\right\}$. Since $\eta_k=\frac{4\mu^{-1}}{kE+\gamma_0}$, it follows that the constraints $\eta_k \leq \frac{1}{\mu}$, $\eta_k \leq \frac{\mu}{L^2}$, and $2L^2(E-1)^2\eta_k^2 \leq 1$ are satisfied.

We now show by induction that for all $k \geq k_0$ which satisfies (30),

$$\delta_k \leq \frac{(k_0+\gamma)^2}{(k+\gamma)^2}\delta_{k_0}+\frac{\alpha}{k+\gamma},\tag{34}$$

which is also used in [25]. Obviously, (34) holds for $k=k_0$ since $\alpha>0$. Assume (34) holds for some $k>k_0$, then

$$\begin{aligned}\delta_{k+1} \leq&\left(1-\frac{2}{k+\gamma}\right)\delta_k+\frac{\alpha}{(k+\gamma)^2}\\\leq&\left(1-\frac{2}{k+\gamma}\right)\left(\frac{(k_0+\gamma)^2}{(k+\gamma)^2}\delta_{k_0}+\frac{\alpha}{k+\gamma}\right)+\frac{\alpha}{(k+\gamma)^2}\\\leq&\frac{(k+\gamma-1)(k_0+\gamma)^2}{(k+\gamma)^3}\delta_{k_0}+\frac{k+\gamma-1}{(k+\gamma)^2}\alpha\\\leq&\frac{(k_0+\gamma)^2}{(k+1+\gamma)^2}\delta_{k_0}+\frac{\alpha}{k+1+\gamma}.\end{aligned}\tag{35}$$

Combining (31) and (34), it follows that

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2 \leq&\frac{(k_0E+\gamma_0)^2}{(kE+\gamma_0)^2}\mathbb{E}\|\mathbf{x}_{k_0}-\mathbf{x}^*\|^2\\&+\tilde{D}_1\frac{E}{kE+\gamma_0}\\&+\tilde{D}_2\frac{1}{kE+\gamma_0}\\&+\tilde{D}_3\frac{E-1}{kE+\gamma_0}\\&+\tilde{D}_4\frac{(E-1)^2}{kE+\gamma_0}\\&+\tilde{D}_5\frac{E^3}{kE+\gamma_0},\end{aligned}\tag{36}$$

where $\tilde{D}_1=\frac{16}{\mu^2}D_1$, $\tilde{D}_2=\frac{16}{\mu^2}D_2$, $\tilde{D}_3=\frac{16}{\mu^2}D_3$, $\tilde{D}_4=\frac{16}{\mu^2}D_4$, and $\tilde{D}_5=\frac{16}{\mu^2}D_5$.

We also observe that the algorithm does not require a warming-up period of more than k_0 iterations as

$$\begin{aligned}\eta_k=\frac{4\mu^{-1}}{kE+\gamma_0} \leq&\frac{1}{\mu E}\\(kE+\gamma_0)\frac{1}{E} \geq&4\\\Rightarrow k \geq 4-\frac{\gamma_0}{E} \Rightarrow&k_0=0\end{aligned}\tag{37}$$

This concludes the proof of Theorem 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

B. Proof of Lemma 1

By denoting $\overline{\Delta}_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\Delta_{k+1}^{(i)}$ and $\Delta_{k+1}^{(i)} =$, the term $\mathbb{E}\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\|^2$ can be calculated as

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\|^2 \\
=&\mathbb{E}\left[\mathbb{E}_{\mathcal{S}_k}\left\|\left(\mathbf{x}_k + \frac{1}{M}\sum_{i\in\mathcal{S}_k}\Delta_{k+1}^{(i)}\right) - \left(\mathbf{x}_k + \overline{\Delta}_{k+1}\right)\right\|^2\right] \\
=&\mathbb{E}\left[\mathbb{E}_{\mathcal{S}_k}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_k}\left(\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\right)\right\|^2\right] \\
=&\frac{1}{M^2}\sum_{i=1}^{N}\Pr\{i\in\mathcal{S}_k\}\mathbb{E}\left\|\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\right\|^2 \\
&+ \frac{1}{M^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\Pr\{i,j\in\mathcal{S}_k\} \\
&\quad\mathbb{E}\left\langle\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}, \Delta_{k+1}^{(j)} - \overline{\Delta}_{k+1}\right\rangle
\end{aligned} \tag{38}$$

$$\begin{aligned}
\overset{(a)}{=}&\frac{1}{MN}\sum_{i=1}^{N}\mathbb{E}\|\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\|^2 \\
&+ \frac{M-1}{MN(N-1)} \\
&\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\mathbb{E}\left\langle\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}, \Delta_{k+1}^{(j)} - \overline{\Delta}_{k+1}\right\rangle \\
\overset{(b)}{=}&\frac{N-M}{M(N-1)}\cdot\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\right\|^2}_{A_1},
\end{aligned} \tag{39}$$

where we use the following facts: (a) $\Pr\{i\in\mathcal{S}_k\} = \frac{M}{N}$ and $\Pr\{i,j\in\mathcal{S}_k\} = \frac{M(M-1)}{N(N-1)}$ for uniform sampling without replacement, and (b) $\sum_{i=1}^{N}\left\|\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\right\|^2 + \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left\langle\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}, \Delta_{k+1}^{(j)} - \overline{\Delta}_{k+1}\right\rangle = 0$.

Term A_1 in Eq. (39) can be further bounded as

$$\begin{aligned}
A_1 =&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\Delta_{k+1}^{(i)} - \overline{\Delta}_{k+1}\right\|^2 \\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\left(\Delta_{k+1}^{(i)} - \Delta_{k+1}^{(j)}\right)\right\|^2
\end{aligned} \tag{40}$$

Then,

$$\begin{aligned}
A_1 \leq&\frac{1}{N^2}\sum_{i\neq j}\mathbb{E}\left\|\Delta_{k+1}^{(i)} - \Delta_{k+1}^{(j)}\right\|^2 \\
=&\frac{1}{N^2}\sum_{i\neq j}\left(\mathbb{E}\left\|\Delta_{k+1}^{(i)}\right\|^2 - 2\mathbb{E}\left\langle\Delta_{k+1}^{(i)}, \Delta_{k+1}^{(j)}\right\rangle + \mathbb{E}\left\|\Delta_{k+1}^{(j)}\right\|^2\right) \\
\leq&\frac{q}{N^2}\sum_{i\neq j}\left(\mathbb{E}\left\|\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k\right\|^2 + \mathbb{E}\left\|\mathbf{x}_{k,E}^{(j)} - \mathbf{x}_k\right\|^2\right) \\
&+ (1+q)\left(\mathbb{E}\left\|\mathbf{w}_k^{(i)}\right\|^2 + \mathbb{E}\left\|\mathbf{w}_k^{(j)}\right\|^2\right) \\
&+ \frac{1}{N^2}\sum_{i\neq j}(\mathbb{E}\left\|\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k\right\|^2 - 2\mathbb{E}\left\langle\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k, \mathbf{x}_{k,E}^{(j)} - \mathbf{x}_k\right\rangle \\
&\quad+ \mathbb{E}\left\|\mathbf{x}_{k,E}^{(j)} - \mathbf{x}_k\right\|^2) \\
=&\frac{q}{N^2}\sum_{i\neq j}\left(\underbrace{\mathbb{E}\left\|\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_k\right\|^2}_{A_2} + \mathbb{E}\left\|\mathbf{x}_{k,E}^{(j)} - \mathbf{x}_k\right\|^2\right) \\
&+ \frac{1}{N^2}\sum_{i\neq j}\underbrace{\mathbb{E}\left\|\mathbf{x}_{k,E}^{(i)} - \mathbf{x}_{k,E}^{(j)}\right\|^2}_{A_3} + d(1+q)\left(\varsigma_{i,k}^2 + \varsigma_{j,k}^2\right),
\end{aligned} \tag{41}$$

where we use Assumption 4 to get the second inequality. A_2 measures the deviation of the local model from $\mathbf{x}_k$ after E local SGD steps, and A_3 measures the difference between the local models on two different devices.

We now bound A_2. Based on Algorithm 2, $\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \mathbf{x}_k\right\|^2$ can be calculated as

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \mathbf{x}_k\right\|^2 =& \eta_k^2\mathbb{E}\left\|\sum_{s=0}^{t-1}\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}; \xi_{k,s}^{(i)}\right)\right\|^2 \\
\overset{(a)}{\leq}& t\eta_k^2\sum_{s=0}^{t-1}\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}; \xi_{k,s}^{(i)}\right)\right\|^2 \\
\overset{(b)}{\leq}& t^2\eta_k^2 G^2,
\end{aligned} \tag{42}$$

where (a) follows by Assumption 3 and the inequality $\left\|\sum_{k=1}^{K}\mathbf{x}_k\right\|^2 \leq K\sum_{k=1}^{K}\|\mathbf{x}_k\|^2$, (b) follows since $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}; \xi_{k,s}^{(i)}\right)\right\|^2 \leq G^2$ due to the Assumption of bounded stochastic gradients.

To bound A_3, we introduce an auxiliary sequence of models $\{\mathbf{a}_{k,0}, \mathbf{a}_{k,1}, \cdots, \mathbf{a}_{k,E}\}$, which is generated by gradient descent on the global loss function, i.e.,

$$\mathbf{a}_{k,t+1} = \mathbf{a}_{k,t} - \eta_k\nabla f(\mathbf{a}_{k,t}), \quad \text{for } t = 0, 1, \cdots, E-1, \tag{43}$$

with $\mathbf{a}_{k,0} = \mathbf{x}_k$. When $i \neq j$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \mathbf{x}_{k,t}^{(j)}\right\|^2 =& \mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \mathbf{a}_{k,t} + \mathbf{a}_{k,t} - \mathbf{x}_{k,t}^{(j)}\right\|^2 \\
\leq& 2\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \mathbf{a}_{k,t}\right\|^2 + 2\mathbb{E}\left\|\mathbf{x}_{k,t}^{(j)} - \mathbf{a}_{k,t}\right\|^2,
\end{aligned} \tag{44}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3101991, IEEE Internet of Things Journal

where

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\mathbf{a}_{k,t}\right\|^2 = \\
&\quad\mathbb{E}\left\|\mathbf{x}_k-\eta_k\sum_{s=0}^{t-1}\nabla f_i\left(\mathbf{x}_{k,s}^{(i)};\xi_{k,s}\right)-\mathbf{x}_k+\eta_k\sum_{s=0}^{t-1}\nabla f\left(\mathbf{a}_{k,s}\right)\right\|^2 \\
&\quad=\eta_k^2\mathbb{E}\left\|\sum_{s=0}^{t-1}\left(\nabla f_i\left(\mathbf{x}_{k,s}^{(i)};\xi_{k,s}\right)-\nabla f(\mathbf{a}_{k,s})\right)\right\|^2 \\
&\quad\overset{(a)}{\leq} t\eta_k^2\sigma^2/b+\eta_k^2\mathbb{E}\left\|\sum_{s=1}^{t-1}\left(\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}\right)-\nabla f(\mathbf{a}_{k,s})\right)\right\|^2 \\
&\quad\leq t\eta_k^2\sigma^2/b+(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}\right)-\nabla f(\mathbf{a}_{k,s})\right\|^2 \\
&\quad=t\eta_k^2\sigma^2/b+(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}\right) \\
&\qquad-\nabla f(\mathbf{a}_{k,s})+\nabla f_i(\mathbf{a}_{k,s})-\nabla f_i(\mathbf{a}_{k,s})\|^2 \\
&\quad\leq t\eta_k^2\sigma^2/b+2(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\nabla f_i(\mathbf{a}_{k,s})-\nabla f(\mathbf{a}_{k,s})\right\|^2 \\
&\quad+2(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}\right)-\nabla f_i(\mathbf{a}_{k,s})\right\|^2 \\
&\quad\overset{(b)}{\leq} t\eta_k^2\sigma^2/b+2(t-1)t\eta_k^2\lambda^2+ \\
&\qquad+2(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_{k,s}^{(i)}\right)-\nabla f_i(\mathbf{a}_{k,s})\right\|^2 \\
&\quad\overset{(c)}{\leq} t\eta_k^2\sigma^2/b+2(t-1)t\eta_k^2\lambda^2+ \\
&\qquad+2L^2(t-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\mathbf{x}_{k,s}^{(i)}-\mathbf{a}_{k,s}\right\|^2 \\
&\quad\overset{(d)}{\leq}\frac{\sigma^2}{b}E\eta_k^2+2\lambda^2(E-1)E\eta_k^2 \\
&\qquad+2L^2(E-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\mathbf{x}_{k,s}^{(i)}-\mathbf{a}_{k,s}\right\|^2, \qquad (45)
\end{aligned}$$

where (a) follows by the bounded variance of stochastic gradient in Assumption 3, (b) follows due to Assumption 5, (c) follows by the L-smoothness, and (d) follows since $t\leq E$ holds.

Now we show by induction that for all $t\leq E$,

$$\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\mathbf{a}_{k,t}\right\|^2\leq\left(E\eta_k^2\sigma^2/b+2\lambda^2E(E-1)\eta_k^2\right)\times\left(1+2L^2(E-1)\eta_k^2\right)^{t-1}. \qquad (46)$$

When $t=1$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k,1}^{(i)}-\mathbf{a}_{k,1}\right\|^2&=\mathbb{E}\left\|\mathbf{x}_k-\eta_k\nabla f_i(\mathbf{x}_k)-\mathbf{x}_k+\eta_k\nabla f(\mathbf{x}_k)\right\|^2 \\
&=\eta_k^2\mathbb{E}\left\|\nabla f_i(\mathbf{x}_k)-\nabla f(\mathbf{x}_k)\right\|^2\leq\eta_k^2\lambda^2, \qquad (47)
\end{aligned}$$

and hence, (46) holds for $t=1$. Assume that (46) holds up to iteration $t-1$, then for iteration t

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\mathbf{a}_{k,t}\right\|^2&\leq\frac{\sigma^2}{b}E\eta_k^2+2\lambda^2(E-1)E\eta_k^2 \\
&\quad+2L^2(E-1)\eta_k^2\sum_{s=1}^{t-1}\mathbb{E}\left\|\mathbf{x}_{k,s}^{(i)}-\mathbf{a}_{k,s}\right\|^2 \\
&\leq\left(\frac{\sigma^2}{b}E\eta_k^2+2\lambda^2(E-1)E\eta_k^2\right)\times \\
&\quad\left(1+2L^2(E-1)\eta_k^2\sum_{s=1}^{t-1}(1+2L^2(E-1)\eta_k^2)^{s-1}\right) \\
&\leq\left(\frac{\sigma^2}{b}E\eta_k^2+2\lambda^2(E-1)E\eta_k^2\right)\times \\
&\quad\left(1+2L^2(E-1)\eta_k^2\frac{(1+2L^2(E-1)\eta_k^2)^{t-1}-1}{2L^2(E-1)\eta_k^2}\right) \\
&\leq\left(\frac{\sigma^2}{b}E\eta_k^2+2\lambda^2E(E-1)\eta_k^2\right)\left(1+2L^2(E-1)\eta_k^2\right)^{t-1}. \qquad (48)
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\mathbf{a}_{k,t}\right\|^2&\leq\left(E\eta_k^2\sigma^2/b+2\lambda^2E(E-1)\eta_k^2\right)\times \\
&\quad\left(1+2L^2(E-1)\eta_k^2\right)^{E-1} \\
&\overset{(a)}{\leq}\left(E\eta_k^2\sigma^2/b+2\lambda^2E(E-1)\eta_k^2\right)\times \\
&\quad e^{2L^2(E-1)^2\eta_k^2} \\
&\overset{(b)}{\leq}\frac{e\sigma^2}{b}E\eta_k^2+2e\lambda^2E(E-1)\eta_k^2, \qquad (49)
\end{aligned}$$

where (a) follows since $1+x\leq e^x$ and (b) follows by assuming that $2L^2(E-1)^2\eta_k^2\leq 1$.

By substituting (49) into (44), A_3 in Eq. (40) is bounded by

$$\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\mathbf{x}_{k,t}^{(j)}\right\|^2\leq\frac{4e\sigma^2}{b}E\eta_k^2+8e\lambda^2E(E-1)\eta_k^2. \qquad (50)$$

Finally, substituting(40), (42), and (50) into (39), we obtain

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1}\right\|^2\leq&\frac{2qG^2(N-M)}{MN}E^2\eta_k^2 \\
&+\frac{4e\sigma^2(N-M)}{bMN}E\eta_k^2 \\
&+\frac{8e\lambda^2(N-M)}{MN}E(E-1)\eta_k^2 \\
&+\frac{2d(1+q)(N-M)}{M(N-1)}\cdot\frac{1}{N}\sum_{i=1}^{N}\varsigma_{i,k}^2.
\end{aligned}$$

By substituting

$$\varsigma_{i,k}^2=\frac{128G^2b^2\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{|\mathcal{D}_i|^2\epsilon^2}E^4\eta_k^2, \qquad (51)$$

we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1}\right\|^2 \leq & \frac{2qG^2(N-M)}{MN}E^2\eta_k^2 \\ & + \frac{4e\sigma^2(N-M)}{bMN}E\eta_k^2 \\ & + \frac{8e\lambda^2(N-M)}{MN}E(E-1)\eta_k^2 \\ & + \frac{256dG^2b^2(1+q)(N-M)\ln\left(\frac{1.25Eb}{|\mathcal{D}_i|\delta}\right)}{M(N-1)|\mathcal{D}_i|^2\epsilon^2}E^4\eta_k^2.\end{aligned}$$

which concludes the proof of Lemma 1.

C. Proof of Lemma 2

For the term B, we have

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}_{k+1}-\overline{\mathbf{x}}_{k,E}\|^2 \\ &= \mathbb{E}\left\|\mathbf{x}_k+\frac{1}{N}\sum_{i=1}^N Q\left(\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right)-\frac{1}{N}\sum_{i=1}^N\mathbf{x}_{k,E}^{(i)}\right\|^2 \\ &= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N Q\left(\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right)-\frac{1}{N}\sum_{i=1}^N\left(\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right)\right\|^2 \\ &\overset{(a)}{=} \frac{1}{N^2}\sum_{i=1}^N\mathbb{E}\left\|Q\left(\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right)-\left(\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right)\right\|^2 \\ &\overset{(b)}{\leq} \frac{q}{N}\cdot\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\mathbf{x}_{k,E}^{(i)}-\mathbf{x}_k\right\|^2,\end{aligned} \tag{52}$$

where (a) and (b) hold by Assumption 4.

Once again, we use the results in (42) and obtain the bound for term B in Eq. (22) as

$$\mathbb{E}\|\hat{\mathbf{x}}_{k+1}-\overline{\mathbf{x}}_{k,E}\|^2 \leq \frac{qG^2}{N}E^2\eta_k^2, \tag{53}$$

which completes the proof for Lemma 2.

D. Proof of Lemma 3

Consider the k-th round of training with the sequence $\{\overline{\mathbf{x}}_{k,0},\overline{\mathbf{x}}_{k,1},\cdots,\overline{\mathbf{x}}_{k,E}\}$ representing the averaged model of all local devices in distributed SGD with $\overline{\mathbf{x}}_{k,0}=\mathbf{x}_k$. A similar problem has been investigated in [23], in which the authors proved the result of one-step distributed SGD as

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t+1}-\mathbf{x}^*\right\|^2 \leq & (1-\mu\eta_k)\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t}-\mathbf{x}^*\right\|^2 \\ & + \underbrace{6L\eta_k^2\Gamma}_{C_1} + \underbrace{\eta_k^2\mathbb{E}\left\|\mathbf{g}_{k,t}-\overline{\mathbf{g}}_{k,t}\right\|^2}_{C_2} \\ & + \underbrace{\frac{2}{N}\sum_{i=1}^N\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\overline{\mathbf{x}}_{k,t}\right\|^2}_{C_3}\end{aligned} \tag{54}$$

with a different measure of heterogeneity, $\Gamma = f(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^N f_i\left(\mathbf{x}_i^*\right)$. Additionally, $\mathbf{g}_{k,t}=\frac{1}{N}\sum_{i=1}^N\nabla f_i(\mathbf{x}_{k,t}^{(i)};\xi_{k,t}^{(i)})$ and $\overline{\mathbf{g}}_{k,t}=\frac{1}{N}\sum_{i=1}^N\nabla f_i(\mathbf{x}_{k,t}^{(i)})$. Since $\mathbf{x}_i^*$ is the optimal model of the i-th local loss function, we have

$$\begin{aligned}f_i(\mathbf{x}^*) &\overset{(a)}{\leq} f_i(\mathbf{x}_i^*)+\frac{1}{2\mu}\left\|\nabla f_i(\mathbf{x}^*)-\nabla f_i(\mathbf{x}_i^*)\right\|^2 \\ &\overset{(b)}{=} f_i\left(\mathbf{x}_i^*\right)+\frac{1}{2\mu}\left\|\nabla f_i(\mathbf{x}^*)-\nabla f(\mathbf{x}^*)\right\|^2 \\ &\overset{(c)}{\leq} f_i\left(\mathbf{x}_i^*\right)+\frac{\lambda^2}{2\mu},\end{aligned} \tag{55}$$

where (a) holds since $f_i(\cdot)$ is μ-strongly convex, (b) holds since $\nabla f_i(\mathbf{x}_i^*)=\nabla f(\mathbf{x}^*)=0$, and (c) holds due to Assumption 5 on heterogeneity. Then, the term C_1 in Eq. (54) can be bounded as

$$\begin{aligned}6L\eta_k^2\Gamma &= 6L\eta_k^2\left(f(\mathbf{x}^*)-\frac{1}{N}\sum_{i=1}^N f_i\left(\mathbf{x}_i^*\right)\right) \\ &= \frac{6L\eta_k^2}{N}\sum_{i=1}^N\left(f_i(\mathbf{x}^*)-f_i\left(\mathbf{x}_i^*\right)\right) \\ &\leq \frac{3L\eta_k^2\lambda^2}{\mu}.\end{aligned}$$

Term C_2 in Eq. (54) can be calculated as

$$\begin{aligned}&\eta_k^2\mathbb{E}\left\|\mathbf{g}_{k,t}-\overline{\mathbf{g}}_{k,t}\right\|^2 \\ &= \eta_k^2\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla f_i(\mathbf{x}_{k,t}^i,\xi_{k,t}^{(i)})-\nabla f_i(\mathbf{x}_{k,t}^i)\right)\right\|^2 \\ &\overset{(a)}{=} \frac{\eta_k^2}{N^2}\sum_{i=1}^N\mathbb{E}\left\|\nabla f_i(\mathbf{x}_{k,t}^i,\xi_{k,t}^{(i)})-\nabla f_i(\mathbf{x}_{k,t}^i)\right\|^2 \\ &\overset{(b)}{\leq} \frac{\eta_k^2\sigma^2}{bN},\end{aligned} \tag{56}$$

where (a) and (b) holds due to Assumption 3.

Finally, we bound term C_3 using the same technique in [23]:

$$\frac{2}{N}\sum_{i=1}^N\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\overline{\mathbf{x}}_{k,t}\right\|^2 \leq 8G^2(E-1)^2\eta_k^2. \tag{57}$$

where (a) follows by the convexity of $\|\cdot\|^2$ and (b) follows by applying Eq. (50).

By substituting (56), (56), and (57) into Eq. (54), we have

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t+1}-\mathbf{x}^*\right\|^2 \leq & (1-\mu\eta_k)\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t}-\mathbf{x}^*\right\|^2 \\ & + \left(\frac{3L\lambda^2}{\mu}+\frac{\sigma^2}{bN}\right)\eta_k^2 + 8G^2(E-1)^2\eta_k^2.\end{aligned}$$

By unfolding the recursions in (58) and since $\overline{\mathbf{x}}_{k,0}=\mathbf{x}_k$,

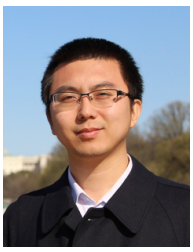
$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,E}-\mathbf{x}^*\right\|^2 \leq & (1-\mu\eta_k)^E\mathbb{E}\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2 \\ & + \left(\frac{3L\lambda^2}{\mu}+\frac{\sigma^2}{bN}\right)\eta_k\frac{1-(1-\mu\eta_k)^E}{\mu} \\ & + 8G^2(E-1)^2\eta_k\frac{1-(1-\mu\eta_k)^E}{\mu} \\ \leq & (1-\mu\eta_k)^E\mathbb{E}\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2 \\ & + \left(\frac{3L\lambda^2}{\mu}+\frac{\sigma^2}{bN}\right)E\eta_k^2 + 8G^2E(E-1)^2\eta_k^2,\end{aligned}$$

where the last inequality holds by using Bernoulli's inequality and the assumption that $\eta_k \leq \frac{1}{\mu}$. To be specific, we have $(1-\mu\eta_k)^s \geq 1 - s\mu\eta_k$, which implies

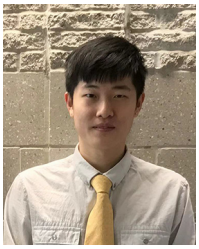
$$\frac{1-(1-\mu\eta_k)^s}{\mu} \leq s\eta_k. \tag{58}$$

Nima Mohammadi (Student Member, IEEE) received his M.Sc degree in Computer Science from the University of Tehran in 2017. He is currently a Ph.D. student at the Bradley Department of Electrical and Computer Engineering at Virginia Tech. His research interests include computational neuroscience and emerging applications of machine learning and neural networks in wireless systems.

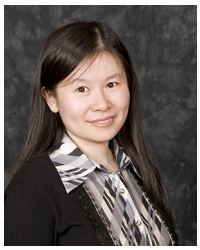
Jianan Bai received the B.S. degree in Communications Engineering from Yingcai Honors College at University of Electronic Science and Technology of China (UESTC) in 2018 and M.S. degree in Electrical Engineering from the Bradley Department of Electrical and Computer Engineering (ECE) at Virginia Tech in 2020, respectively. His research interest is applying tools from optimization, stochastic geometry, and machine learning to solve emerging problems in device-to-device (D2D), internet-of-things (IoT), and machine-type-communications (MTC) networks.



Qiang Fan (Member, IEEE) received his Ph.D. degree in Electrical and Computer Engineering from New Jersey Institute of Technology (NJIT) in 2019, and his M.S. degree in Electrical Engineering from Yunnan University of Nationalities, China, in 2013. He was a postdoctor researcher in the Department of Electrical and Computer Engineering, Virginia Tech. He has served as a reviewer for over 120 journal submissions such as IEEE Transactions on Cloud Computing, IEEE Journal on Selected Areas in Communications, IEEE Transactions on Communications. His current research interests include wireless communications and networking, mobile edge computing, machine learning and drone assisted networking.



Yifei Song received his B.S. degree in Electrical Engineering and M.S. in Electrical and Computer Engineering from University of Connecticut and University of Washington respectively in 2017 and 2020. He joined the department of Electrical and Computer Engineering at Virginia Tech as a Ph.D. student in 2021 Spring. His current research interests are in the broad area of wireless communications, machine learning and optimization.



Yang Yi (Senior Member, IEEE) is currently an Associate Professor with The Bradley Department of Electrical Engineering and Computer Engineering, Virginia Tech. Her research interests include very large scale integrated (VLSI) circuits and systems, computer-aided design (CAD), neuromorphic architecture for brain-inspired computing systems, and low-power circuits design with advanced nanotechnologies for high-speed wireless systems.

Lingjia Liu (Senior Member, IEEE) received his B.S. degree in Electronic Engineering from Shanghai Jiao Tong University and Ph.D. degree in Electrical and Computer Engineering from Texas A&M University. Prior to joining the ECE Department at Virginia Tech (VT), he was an Associate Professor in the EECS Department at the University of Kansas (KU). He spent 4+ years working in Mitsubishi Electric Research Laboratory (MERL) and the Standards & Mobility Innovation Lab of Samsung Research America (SRA) where he received Global Samsung Best Paper Award in 2008 and 2010. He was leading Samsung's efforts on multiuser MIMO, CoMP, and HetNets in 3GPP LTE/LTE-Advanced standards. Dr. Liu's general research interests mainly lie in emerging technologies for Beyond 5G cellular networks including machine learning for wireless networks, massive MIMO, massive MTC communications, and mmWave communications.