Differentially Private Federated Clustering Over Non-IID Data

Yiwei Li, *Member, IEEE*, Shuai Wang, Chong-Yung Chi, *Life Fellow, IEEE*, and Tony Q. S. Quek, *Fellow, IEEE*

***Abstract*—In this article, we investigate the federated clustering (FedC) problem, which aims to accurately partition unlabeled data samples distributed over massive clients into finite clusters under the orchestration of a parameter server (PS), meanwhile considering data privacy. Though it is an NP-hard optimization problem involving real variables denoting cluster centroids and binary variables denoting the cluster membership of each data sample, we judiciously reformulate the FedC problem into a non-convex optimization problem with only one convex constraint, accordingly yielding a soft clustering solution. Then, a novel FedC algorithm using differential privacy (DP) technique, referred to as DP-FedC, is proposed in which partial clients participation (PCP) and multiple local model updating steps are also considered. Furthermore, various attributes of the proposed DP-FedC are obtained through theoretical analyses of privacy protection and convergence rate, especially for the case of nonidentically and independently distributed (non-i.i.d.) data, that ideally serve as the guidelines for the design of the proposed DP-FedC. Then, some experimental results on two real datasets are provided to demonstrate the efficacy of the proposed DP-FedC together with its much superior performance over some state-of-the-art FedC algorithms, and the consistency with all the presented analytical results.**

***Index Terms*—Differential privacy (DP), federated clustering (FedC), non-i.i.d. data, privacy amplification.**

I. Introduction

FEDERATED learning (FL) , as a novel distributed paradigm, enables massively distributed clients to jointly find a desired model through machine learning (ML) under the orchestration of a parameter server (PS) while refraining the clients' sensitive data from being exposed [1], [2]. FL has received tremendous attention in the past several years as it seriously takes numerous practical challenges into account, including limited communication resources and data heterogeneity and client privacy protection in the learning process [3]. Under these challenges, most FL algorithms follow a computation-aggregation protocol by which the local update of model parameters and the PS aggregation are repeated in a round-by-round fashion until convergence. Federated average (FedAvg) algorithm [4], [5] is a typical one, which improves communication efficiency by adopting partial client participation (PCP) and multiple local stochastic gradient descent (local SGD) updating steps. Nevertheless, the data heterogeneity (e.g., nonidentically and independently distributed (non-i.i.d.) data) has been acknowledged to be the main bottleneck to FL deployment. Numerous efforts have been devoted to the analysis of the adverse effects of non-i.i.d. data on algorithm convergence [4], [6]. In parallel, FL still suffers from privacy leakage as the clients' sensitive information could be inferred by adversaries through the exchanged model parameters between the clients and the PS [7], [8], [9]. As illustrated in Fig. 1, a vanilla FL system includes many clients and one PS, where the uploaded parameters from local clients may be overheard by powerful adversaries. The differential privacy (DP) technique has recently gained increasing popularity in enhancing privacy of FL thanks to its algorithmic simplicity, support by rigorous mathematical theory, and negligible system overheads [10], [11].

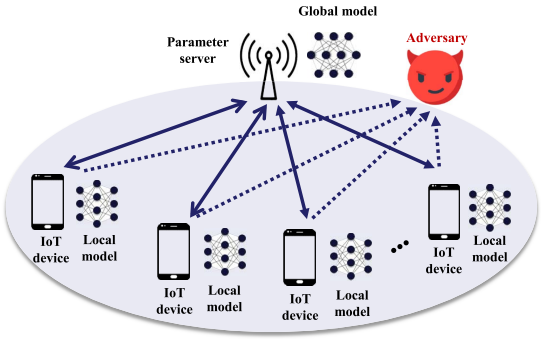


Fig. 1. Framework of the FL system in the presence of adversaries.

Despite the recent rapid progress of FL, substantial attention has been given to supervised learning, whereas the problem of federated unsupervised learning, especially data clustering, has not yet been investigated comprehensively in FL community [12]. Clustering in the FL setting, called federated clustering (FedC), aims to partition data samples

Manuscript received 2 January 2023; revised 30 July 2023; accepted 1 September 2023. Date of publication 7 September 2023; date of current version 6 February 2024. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 111-2221-E-007-035-MY2 and Grant MOST 110-2221-E-007-031, and in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme. *(Corresponding authors: Shuai Wang; Tony Q. S. Quek.)*

Yiwei Li and Chong-Yung Chi are with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: lywei0306@foxmail.com; cychi@ee.nthu.edu.tw).

Shuai Wang is with the Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372 (e-mail: shuaiwang@link.cuhk.edu.cn).

Tony Q. S. Quek is with the Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372, and also with the Yonsei Frontier Laboratory, Yonsei University, Seoul 03722, South Korea (e-mail: tonyquek@sutd.edu.sg).

Digital Object Identifier 10.1109/JIOT.2023.3312852

distributed over massive clients based on a global similarity measure while keeping them on respective clients. As clustering is one of most suitable missions for ML and has a great deal of applications, the FedC and its implementation is believed to be in impending need. On the other hand, recent years have witnessed an incessant springing up of FedC applications, which again motivates research efforts in this direction. For example, in e-commerce applications, FedC is widely used to group the online customers of multiple institutions with sensitive features, such as personal details, purchase orders, and bank transaction records, to identify their specific interests for precise service recommendation [13], [14]. Note that, FedC is quite different from clustered FL approaches [15], [16], which, instead of data clustering, are concerned with clusters of clients such that each cluster comes up with a local model to be uploaded to the PS in order to reduce the communication cost of supervised FL systems [17], [18], [19].

In this article, an effective FedC algorithm is proposed, that considers both non-i.i.d. data and DP-based privacy protection. In FedC scenarios where data heterogeneity is prevalent, the global cluster information may not be available for each client as all the data in hands may belong to just a few clusters, and the correct cluster structure might become apparent only when the local datasets are combined [19]. Moreover, effectively transferring the centralized clustering algorithms into FedC, such as k-means, is almost formidable due to the privacy concern. Directly applying them to FedC by following the computation-aggregation protocol would result in serious performance degradation [12], [20]. In addition, different from supervised FL, the process of FedC involves the iterative constrained optimization of both cluster centroids and cluster assignments of all data samples, which again brings more difficulties to algorithm design.

As for privacy protection, such coupling optimization necessitates a more careful and fine-grained design and analysis of the DP-based FedC algorithms. In particular, it is widely known that DP protects privacy at the cost of learning performance loss [21], and balancing the tradeoff between protection level and convergence performance, so-called the privacy–utility tradeoff, is essential in practical FL applications. To improve the privacy–utility tradeoff, privacy amplification [22], [23] has been pervasively adopted in many DP-based FL (DP-FL) applications [5], [24]. The privacy amplification can reduce the variance of noise added to locally uploaded models without sacrificing the privacy protection level, thereby mitigating the adverse effects of DP [25]. In addition to the challenges posed by non-i.i.d. data and privacy protection, the practical application of FedC algorithms in FL systems requires careful consideration of communication cost and straggler effect [26]. These factors and concerns not only affect the algorithm design but also make the associated theoretical algorithm performance analysis much more involved. However, the involvement of cluster centroids and data's cluster-membership assignment in FedC further complicates the design of DP, and it is still not clear how to achieve a good privacy–utility tradeoff in FedC.

A. Related Works

Currently, many successful methods have been reported about traditional distributed clustering, however, they are simply parallel implementations of the centralized clustering algorithms [27], [28], [29] or implementations through clustering representative data samples collected from distributed clients [20], [30]. Apparently, the critical challenges of FL, such as massive clients, limited communication resources and data heterogeneity, were rarely considered, and the demand for privacy protection was also overlooked.

The recent works in [13], [31], [32], and [33] have considered the FL scenarios and presented FedC algorithms, while most of them were developed by combining the simple centralized k-means algorithm (and its variants) with FedAvg [4]. Specifically, in each communication round, the clients employ k-means algorithms to obtain the local cluster centroids, which are then uploaded to the PS to produce the global clusters. The works [13], [31] adopted the fuzzy k-means to perform local clustering, while the global centroids are obtained from the received local centroids by k-means clustering. The work [32] proposed a federated spectral clustering approach to train a generative model for each cluster, such that each data sample can be classified to only one cluster using the generated models. Nevertheless, the above-mentioned FedC algorithms did not consider the data heterogeneity issue; thus, hardly yielding satisfactory clustering performance as the clustering algorithm only works well in clustering datasets that are evenly spread around the centroids but fails in clustering datasets of complex and heterogeneous cluster structure [19], [20], [34].

To the best of our knowledge, only few works have specifically addressed FedC in the context of non-i.i.d. data [15], [18], [35], [36]. However, these works also have their limitations. The approach reported in [15] directly apply the conventional k-means to the FL framework, resulting in suboptimal clustering performance, that will be discussed in Section II-B. The work [18] considered one-shot FedC, where each client obtains a local model using k-means and then upload the trained model only once for the aggregation by the PS. However, the one-shot FedC may not be very effective when the FL problem under consideration is NP-hard or nonconvex due to the low-quality of local solutions. The work [36] proposed a FedC scheme by assuming that heterogeneous data to be clustered come from a probabilistic model, that is, assign each data point to a cluster model with the highest likelihood. More importantly, these works [15], [18], [36] lack a complete theoretical analysis of the impact of non-i.i.d. data on convergence performance. The work in [35] formulates the clustering problem as a constrained nonconvex problem and theoretically analyze the impact of non-i.i.d. data. However, none of above-mentioned works ever consider the crucial issue of privacy protection, which we believe, is one of the most fundamental concerns in FL system. The work [5] is the first that adopted a secret sharing approach to protect privacy in the federated k-means algorithm. However, such a strategy requires complicated encryption protocols and substantial extra communication and computation cost [37], thus not applicable to large-scale FL models. As far as we are aware, none of the

existing works simultaneously consider data heterogeneity and privacy protection, hence motivating us to develop advanced privacy-preserving FedC algorithms over non-i.i.d. data.

B. Contributions

Motivated by the aforementioned issues of existing FedC methods, we propose a differentially private FedC algorithm, called DP-FedC, with the data heterogeneity and privacy protection taken into account. The main contributions of this work are summarized as follows.

1) A novel clustering problem is formulated to overcome the shortcomings of the conventional centralized k-means, then applied it to FedC scenarios. To handle the proposed FedC problem, a DP-FedC algorithm under the computation-aggregation protocol is developed, that alternatively update local cluster centroids and indicator matrices (indicating each sample and the cluster it belongs) through allowing multiple local SGD steps and partial clients PCP. Furthermore, the privacy amplification strategy is employed to reduce the DP noise variance for better tradeoff between learning performance (i.e., clustering accuracy) and privacy protection.
2) Two theoretical analyses for the proposed DP-FedC algorithm are presented. One is a privacy analysis, showing that a tighter upper bound of the total privacy loss, i.e., $(\mathcal{O}(q\epsilon\sqrt{pR}), \delta)$-DP over R consecutive communication rounds, where $0 < p, q \leq 1$ are defined in Remark 2. The other is a convergence analysis, showing the convergence rate $\mathcal{O}(1/\sqrt{R})$ under nonconvex and non-i.i.d. data setting.
3) Extensive experimental results are provided to demonstrate the effectiveness of the proposed DP-FedC algorithm on real world datasets, including TCGA cancer gene data [38], and the MNIST handwriting digits data [39], and its much superior performance over state-of-the-art distributed clustering and FedC algorithms.

Synopsis: Section II introduces some preliminaries of DP. Section III presents the problem formulation. Section IV presents the proposed DP-FedC algorithm. Section V focuses on privacy analysis and convergence analysis of the proposed algorithm. Experiment results are presented in Section VI, and finally the conclusion is drawn in Section VII.

Notation: $\mathbb{E}[\cdot]$ represents the expectation of random variables or events; $\Pr[\cdot]$ represents the probability function; $\mathbb{R}^{m\times n}$ denotes the set of m by n real-valued matrices; the (i,j)th entry of matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is denoted by $\mathbf{A}(i,j)$; $\mathbf{A}(i,:)$ and $\mathbf{A}(:,j)$ denote the ith row and the jth column of $\mathbf{A}$, respectively; $\mathbf{A} \geq 0$ means $\mathbf{A}(i,j) \geq 0, \forall i,j$; $[\mathbf{A}]^+$ denotes the matrix by replacing all the negative elements in $\mathbf{A}$ with zero. $\lambda_{\max}(\mathbf{A})$ stands for the maximum eigenvalue of $\mathbf{A}$; $\|\cdot\|_F$, $\|\cdot\|$, and $\|\cdot\|_0$ are the matrix Frobenius norm, Euclidean norm, (i.e., ℓ_2-norm) and zero norm of vectors, respectively; $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^\top\mathbf{y}$ represents the inner product operator, where the superscript '$\top$' denotes the vector transpose; for any integer N, $[N]$ denotes the integer set $\{1,\ldots,N\}$; $\mathbf{1}$ denotes the all-one vector; and $\{\mathcal{C}_i\}_{i=1}^k$ denotes the set $\{\mathcal{C}_1, \mathcal{C}_2 \ldots, \mathcal{C}_k\}$; $\lfloor\cdot\rfloor$ denotes the floor function.

II. Preliminaries

A. Differential Privacy

In this work, we assume that any third party is untrustworthy, including the honest-but-curious server. The core privacy protection mechanism of the proposed DP-FedC is the well-known DP based random mechanism defined as follows.

Definition 1 (ϵ, δ)-DP [11]: Consider two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, which differ in only one data sample. A randomized mechanism $\mathcal{M}$ is (ϵ,δ)-DP if for any two $\mathcal{D}$, $\mathcal{D}'$ and measurable subset $\boldsymbol{O} \subseteq \text{Range}(\mathcal{M})$

$$\Pr[\mathcal{M}(\mathcal{D}) \in \boldsymbol{O}] \leq \exp(\epsilon)\cdot \Pr\big[\mathcal{M}\big(\mathcal{D}'\big) \in \boldsymbol{O}\big] + \delta \tag{1}$$

holds true, $\epsilon > 0$ represents the privacy protection level, and $0 < \delta < 1$ is the probability threshold to break $(\epsilon, 0)$-DP.

A smaller ϵ means that it is more difficult to distinguish between the two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, thus resulting in stronger privacy protection. The required "noise variance" σ^2 for achieving (ϵ,δ)-DP is given by the following lemma.

Lemma 1 [11, Th. 3.22]: Suppose a query function g accesses the dataset $\mathcal{D}$ via randomized mechanism $\mathcal{M}$. Let ξ be a zero-mean Gaussian noise with variance σ^2. Then, $g+\xi$ is (ϵ,δ)-DP if

$$\sigma^2 = \frac{2s^2\ln(1.25/\delta)}{\epsilon^2} \tag{2}$$

where s is the ℓ_2-norm sensitivity of the function g defined by

$$s \triangleq \max_{\mathcal{D},\mathcal{D}'} \big\|g(\mathcal{D}) - g\big(\mathcal{D}'\big)\big\|. \tag{3}$$

In practical FL systems, it is crucial to monitor the total privacy loss over multiple communication rounds of model parameters exchange with the PS, which can be computed from the individual privacy loss stated in the following definition.

Definition 2 (Privacy Loss [11]): Suppose that a randomized mechanism $\mathcal{M}$ satisfies (ϵ,δ)-DP. Let $\mathcal{D}$ and $\mathcal{D}'$ be two neighboring datasets and $\boldsymbol{O}$ be a possible random vector of $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$. Then, the privacy loss is defined by

$$\text{PL}(\boldsymbol{O}) \triangleq \ln\left(\frac{\mathbb{P}\big[\mathcal{M}(\mathcal{D}) = \boldsymbol{O}\big]}{\mathbb{P}\big[\mathcal{M}(\mathcal{D}') = \boldsymbol{O}\big]}\right). \tag{4}$$

Note that, the computation of total privacy loss is quite involved, though its upper bound can be estimated using the moment accountant method [40], which so far yields the tightest bound on the total privacy loss.

According to the privacy amplification theorem [23], it has been known that, running on a randomly generated subset of a dataset, the DP mechanism can yield stronger privacy protection than running on the entire dataset. This fact implies that the noise variance required for achieving a predefined DP level can be reduced when partial data are randomly selected at each iteration. The privacy analysis to be addressed in Section V-B relies on the following privacy amplification theorem.

Theorem 1 (Privacy Amplification Theorem [23]): Suppose that a mechanism $\mathcal{M}$ is (ϵ,δ)-DP over a given dataset $\mathcal{D}$ with size n. Consider the subsampling mechanism that outputs a random sample uniformly over all subsets $\mathcal{D}_s \subseteq \mathcal{D}$ with size b.

Then, when $\epsilon \leq 1$, executing $\mathcal{M}$ mechanism on the subset $\mathcal{D}_s$ guarantees (ϵ', δ')-DP, where ϵ' and δ' are given by

$$\epsilon' = \min(2q\epsilon, \epsilon), \delta' = q\delta \tag{5}$$

where $q = b/n$ is the data sampling ratio.

Proof: See Appendix A. ■

According to Theorem 1, the privacy would be amplified when $q \leq 1/2$. Note that, the privacy amplification for local DP is pervasively adopted in existing FL literatures [10], [24] since only a small portion of data being used in local SGD.

B. Centralized k-Means Clustering

Let $\mathbf{X}$ be a data matrix that contains n data samples and each sample has m features, i.e., $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$. The clustering task is to assign the n data samples of $\mathbf{X}$ to a predefined number of k clusters such that the samples within a cluster are closer to each other than to those belonging to any other cluster in terms of a certain distance metric. Among hundreds of clustering algorithms, the most classic and popular one is the k-means which aims to obtain k nonoverlapping clusters $\{\mathcal{C}_i\}_{i=1}^k$, i.e., $\mathcal{C}_i \cap \mathcal{C}_{i'} = \emptyset \quad \forall i \neq i' \in [k], \bigcup_{i=1}^k \mathcal{C}_i = \{\mathbf{x}_j\}_{j=1}^n$, by minimizing the average Euclidean distance between each cluster centroid and all the data samples within the cluster.

From the optimization perspective, the k-means algorithm can be viewed as an ad hoc algorithm, which handles the following matrix factorization by alternative minimization (AM) [41]:

$$\min_{\mathbf{W}\in\mathbb{R}^{m\times k},\mathbf{H}} \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \tag{6a}$$

$$\text{s.t.} \quad \mathbf{H} \in \{0, 1\}^{k\times n}, \|\mathbf{H}(:, j)\|_0 = 1 \quad \forall j \tag{6b}$$

where $\mathbf{W} \in \mathbb{R}^{m\times k}$ is a matrix consisting of the k centroids, and $\mathbf{H} \in \mathbb{R}^{k\times n}$ is an indicator matrix with only one nonzero element (i.e., unity) in each column. Applying AM to problem (6) gives rise to the following update rules of $\mathbf{W}$ and $\mathbf{H}$ at iteration $t + 1$:

$$\mathbf{H}^{t+1} = \arg\min_{\mathbf{H}} \|\mathbf{X} - \mathbf{W}^t\mathbf{H}\|_F^2$$

$$\text{s.t.} \quad \mathbf{H} \in \{0, 1\}^{k\times n}, \|\mathbf{H}(:, j)\|_0 = 1 \quad \forall j \tag{7}$$

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}\in\mathbb{R}^{m\times k}} \|\mathbf{X} - \mathbf{W}\mathbf{H}^{t+1}\|_F^2. \tag{8}$$

The closed-form solutions to (7) and (8) are, respectively, given by

$$\mathbf{H}^{t+1}(l, j) = \begin{cases} 1, \text{ if } l = \arg\min_u \|\mathbf{X}(:, j) - \mathbf{W}^t(:, u)\|^2 \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

and

$$\mathbf{W}^{t+1}(:, l) = \frac{1}{|\mathcal{J}_l^t|} \sum_{u\in\mathcal{J}_l^t} \mathbf{X}(:, u) \tag{10}$$

where $\mathcal{J}_l^t = \{j|\mathbf{H}(l; j) \ = 1\}$. Note that at iteration $t + 1$, the lth row of $\mathbf{H}$ is updated according to the minimum distance from each data sample to the lth centroid according to $\mathbf{W}^t$, and then the lth centroid (i.e., the l-column of $\mathbf{W}$) is updated as the average of the data belonging to cluster l according to the lth row of the updated indicator matrix.

III. Problem Formulation

The centralized k-means may totally fail for the dataset with complex distribution and data heterogeneity, so it is not suitable for distributed environments, especially the FL setting. The reasons are twofold. First, the nonconvex k-means problem (6) is NP-hard due to involving binary variables, and hence almost any algorithm (including k-means) is unable to work well. No wonder, it is performance is quite sensitive to the initial conditions, complex data distribution, and the obtained solution easily trapped in bad local minima and so forth [42]. Moreover, the less data samples the worse its performance, thus further downgrading its performance in FL scenarios, especially when the data are sensitive and under privacy concern. Most existing FedC algorithms are based on k-means and operate in computation-aggregation fashion, but their performance may get seriously downgraded under FL scenarios, including massively distributed clients and severe client heterogeneity [43].

Inspired by the idea in [44], we replace the binary constraint (6b) with a norm-based equality constraint and reformulate problem (6) as

$$\min_{\mathbf{W},\mathbf{H}} \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \frac{\mu_h}{2}\|\mathbf{H}\|_F^2 + \frac{\mu_w}{2}\|\mathbf{W}\|_F^2 \tag{11a}$$

$$\text{s.t.} \quad \mathbf{H} \geq 0, \|\mathbf{H}(:, j)\|_1^2 = \|\mathbf{H}(:, j)\|_2^2 \quad \forall j \in [n] \tag{11b}$$

where $\mu_h > 0$ and $\mu_w > 0$ are two positive parameters. Problem (11) is a nonconvex and nonsmooth problem and it can be regarded as a relaxation of the k-means problem (6) because $\mathbf{H}$ has been relaxed as a real $k \times n$ matrix, with at most one nonzero entry (not equal to one) in each column, though the equality constraint (11b) is still nonconvex. Moreover, the two regularization terms [i.e., the 2nd and the 3rd terms in (11a)] are used to control the resulting scaling/counter-scaling ambiguity [41].

Instead of directly solving problem (11), we consider the following problem by dropping the equality constraint in (11b) and adding an associated penalty term in the objective function

$$\min_{\mathbf{W},\mathbf{H}} \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \frac{\mu_h}{2}\|\mathbf{H}\|_F^2 + \frac{\mu_w}{2}\|\mathbf{W}\|_F^2$$

$$+ \frac{\rho}{2}\sum_{j=1}^n \left(\left(\mathbf{1}^\top\mathbf{H}(:, j)\right)^2 - \|\mathbf{H}(:, j)\|_2^2\right) \tag{12a}$$

$$\text{s.t.} \quad \mathbf{H} \geq 0 \tag{12b}$$

where $\rho > 0$ is a penalty parameter. The larger the value of ρ, the smaller the approximation error of the equality constraint in (11b) and the more sparse the matrix $\mathbf{H}$. It is remarkable that problem (12) is much efficient to handle than problem (11) for two reasons. One is that (12b) is a simple convex constraint; the other is that the assignment of each data sample to an unique cluster is not reliable for problem (11) [45], [46]. Therefore, in contrast to the hard clustering performed by k-means, solving (12) corresponds to seeking a soft clustering solution [47] instead.

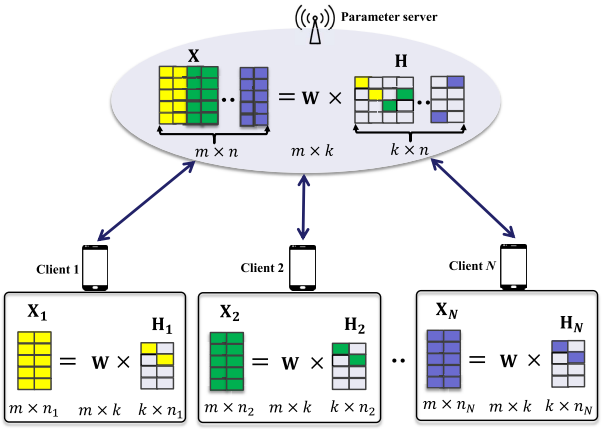


Fig. 2. Proposed framework for FedC.

A. Federated Clustering Model

To solve problem (12) under the FL network, we first assume that the data matrix is partitioned and distributed over N clients. i.e., $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N]$. Specifically, each client i owns nonoverlapping data $\mathbf{X}_i \in \mathbb{R}^{m\times n_i}$, where n_i is the number of data samples in client i and $\sum_{i=1}^{N} n_i = n$. Under the FL scenario, N could be large, and the data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ could be unbalanced and non-i.i.d. [48], [49]. We proceed by partitioning $\mathbf{H}$ in the same fashion as $\mathbf{X}$, resulting in the form $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_N]$. Each column of $\mathbf{H}$ corresponds to a certain data sample in $\mathbf{X}$, while $\mathbf{W}$ is treated as shared parameters that will be uploaded to the PS for information exchange. The resulting framework for FedC is illustrated in Fig. 2, where a central PS coordinates the N clients to accomplish the clustering task. Then, one can reformulate problem (12) into the FL framework as follows:

$$\min_{\substack{\mathbf{W},\ \mathbf{H}_i,\\ i=1,\ldots,N}} \quad F(\mathbf{W},\mathbf{H}) \triangleq \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{W},\mathbf{H}_i) \tag{13a}$$

$$\text{s.t.} \quad \mathbf{H}_i \geq 0 \quad \forall i \in [N] \tag{13b}$$

where

$$F_i(\mathbf{W},\mathbf{H}_i) \triangleq \|\mathbf{X}_i - \mathbf{W}\mathbf{H}_i\|_F^2 + \frac{\rho}{2}\left(\mathrm{Tr}\left(\mathbf{H}_i\mathbf{U}\mathbf{H}_i^\top\right) - \|\mathbf{H}_i\|_F^2\right) + \frac{\mu_h}{2}\|\mathbf{H}_i\|_F^2 + \frac{\mu_w}{2}\|\mathbf{W}\|_F^2 \tag{14}$$

is the local objective function of each client i, and $\mathbf{U} \triangleq \mathbf{1}\mathbf{1}^\top$.

In contrast to the vanilla FL problem which contains only one shared optimization variable, problem (13) involves two variables: one is $\mathbf{W}$ which is the cluster centroid matrix $\mathbf{W}$ shared among clients, and the other one is $\mathbf{H}_i$ which is local cluster indicator matrix for $\mathbf{X}_i$ owned by client i. This apparently brings challenges in the algorithm development, especially in the presence of non-i.i.d. data. In parallel, as $\mathbf{W}$ is shared, there certainly exist possibilities of leaking clients' privacy in the FL process. Recent work [50] showed that the honest-but-curious server could infer clients' private data from the uploaded information in the federated matrix factorization framework. Consequently, it is inevitable to develop an effective and privacy-preserving FL algorithm for problem (13).

IV. Proposed Algorithm for Problem (13)

In this section, we develop a novel FedC algorithm to solve (13), which judiciously updates $\mathbf{W}$ and $\mathbf{H}_i$, $i \in [N]$, and adopts an amplified DP for rigorous privacy protection.

A. Update of $\boldsymbol{W}$ and $\boldsymbol{H}_i$ in FL

The key of algorithmic development to problem (13) is to specify how to perform the local update of $\mathbf{H}_i$ and global update of $\mathbf{W}$. Inspired by [26], we follow the same spirit of local SGD and PCP, where a subset of clients are selected to locally update $\mathbf{H}_i$ and the associated local copies $\mathbf{W}_i$'s of $\mathbf{W}$, and then upload these iterates to the PS for global aggregation in each round. In particular, for round $t = 1, 2, \ldots$

1) *Client Sampling:* We let the PS uniformly sample a small and fixed-size set $\mathcal{S}^t$ of K clients, i.e., $\mathcal{S}^t \subseteq [N]$, $|\mathcal{S}^t| = K$, and then broadcast the global $\mathbf{W}^{t-1}$ to all clients.
2) *Local Update:* All clients are asked to obtain an approximate solution $(\mathbf{W}_i^t, \mathbf{H}_i^t)$ to the following local subproblem of (13)

$$\left(\mathbf{W}_i^t, \mathbf{H}_i^t\right) = \arg\min_{\mathbf{W},\mathbf{H}_i\geq 0} \quad F_i(\mathbf{W},\mathbf{H}_i). \tag{15}$$

After that, each client $i \in \mathcal{S}^t$ uploads $\mathbf{W}_i^t$ to the PS.
3) *Global Aggregation:* After receiving $\mathbf{W}_i^t$ from all clients $i \in \mathcal{S}^t$, the PS aggregates them to produce the new global $\mathbf{W}^t$, i.e.,

$$\mathbf{W}^t = \frac{1}{K}\sum_{i\in\mathcal{S}^t} \mathbf{W}_i^t. \tag{16}$$

In order to specify the local iterates $(\mathbf{W}_i^t, \mathbf{H}_i^t)$, we propose to handle (15) by combining AM [51] and local SGD. That is, $\mathbf{H}_i^t$ is produced by applying multiple gradient descent (GD) steps to (15) with $\mathbf{W}_i$ fixed, and then $\mathbf{W}_i^t$ is updated similarly by fixing $\mathbf{H}_i$. To be more specific, we first let all clients perform $Q_1 \geq 1$ consecutive steps of projected GD with respect to $\mathbf{H}_i$, i.e., for $r = 1, \ldots, Q_1$

$$\mathbf{H}_i^{t,r} = \left[\mathbf{H}_i^{t,r-1} - \tfrac{1}{\gamma_i^t}\nabla_{H_i} F_i\left(\mathbf{W}^{t-1}, \mathbf{H}_i^{t,r-1}\right)\right]^+ \tag{17}$$

where $\gamma_i^t > 0$ is the learning rate. Then, they are asked to perform $Q_2^t \geq 1$ consecutive steps of SGD (no projection) with respect to $\mathbf{W}$, i.e., for $r = Q_1 + 1, \ldots, Q^t$,

$$\mathbf{W}_i^{t,r} = \mathbf{W}_i^{t,r-1} - \tfrac{1}{\eta^t}\nabla_W F_i\left(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^{t,Q_1}; \mathcal{B}_i^{t,r}\right) \tag{18}$$

where $Q^t = Q_1 + Q_2^t$ and $\eta^t > 0$ is a step size, and $\nabla_W F_i(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^t; \mathcal{B}_i^{t,r})$ is the stochastic gradient computed using minibatch dataset $\mathcal{B}_i^{t,r}$ with size b ($|\mathcal{B}_i^{t,r}| = b$). Finally, $(\mathbf{W}_i^t, \mathbf{H}_i^t)$ is obtained by setting $\mathbf{H}_i^t = \mathbf{H}_i^{t,Q_1}$ and $\mathbf{W}_i^t = \mathbf{W}_i^{t,Q^t}$.

B. Privacy Concern

Data privacy is one of primary concerns in FL systems. To enhance data privacy, we apply the DP technique to the proposed algorithm. Specifically, in each round t, we add an artificially Gaussian noise matrix $\boldsymbol{\xi}_i^t \in \mathbb{R}^{m\times k}$ to $\mathbf{W}_i^t$, where all the mk entries of $\boldsymbol{\xi}_i^t$ are i.i.d. Gaussian random variables with zero mean and variance $\sigma_{i,t}^2$, thus yielding

$$\widetilde{\mathbf{W}}_i^t = \mathbf{W}_i^t + \boldsymbol{\xi}_i^t \tag{19}$$

and then upload $\widetilde{\mathbf{W}}_i^t$ to the PS. Then, (16) becomes

$$\mathbf{W}^{t+1} = \frac{1}{K}\sum_{i\in\mathcal{S}^t}\widetilde{\mathbf{W}}_i^t. \tag{20}$$

The details of the proposed algorithm are summarized in Algorithm 1. Note that, the diminishing $Q_2^t = \lfloor(\widehat{Q}/t)\rfloor + 1$ (line 12) denotes the number of iterations in updating $\mathbf{W}_i^{t,r}$ (lines 13–15) by (18), where $\widehat{Q}$ is a given constant and the minibatch dataset $\mathcal{B}_i^{t,r}$ of size b used is further discussed in the following remark.

Remark 1: For lines 13–15 of Algorithm 1, $Q_2^t b$ data samples are obtained from the dataset $\mathcal{D}_i$ at each communication round (i.e., the data sampling ratio $q_{i,t} = Q_2^t b/n_i$), and then divided into Q_2^t minibatch datasets $\mathcal{B}_i^{t,r}$ for each inner iteration r.

It is acknowledged that the DP noise matrix $\boldsymbol{\xi}_i^t$ will bring about adverse effects on algorithm convergence and learning performance. However, the performance degradation of Algorithm 1 will get worse from round to round due to $\mathbf{W}$ perturbed by the DP noise and the coupling of $\mathbf{W}$ and $\mathbf{H}$, on the one hand. The accumulated DP noise effects will also get worse with t on the other hand. Therefore, Algorithm 1 is performance-sensitive to the DP noise in a complicated manner, such that obtaining a satisfactory privacy–utility tradeoff through theoretical analysis becomes more intractable.

Nevertheless, the privacy amplification presented in Theorem 1, can be utilized to pursue the performance analysis of Algorithm 1, in order to find the clue about the variance reduction of the DP noise for guaranteeing (ϵ, δ)-DP privacy protection level at each round. The details are presented in the next section.

Algorithm 1 DP-FedC Algorithm

1: **Input:** initial values of $\mathbf{W}_1^0 = \cdots = \mathbf{W}_N^0 = \mathbf{W}^0$, initial values of $\{\mathbf{H}_i^0\}_{i=1}^N$, $\mathcal{S}^0 = \{1, \ldots, N\}$, R and $\widehat{Q}$.
2: **for** round $t = 1$ **to** R **do**
3: **Server side:**
4: Compute $\mathbf{W}^t$ by (20).
5: Uniformly sample a set of clients $\mathcal{S}^t \subseteq [N]$, and broadcast $\mathbf{W}^t$ to all clients.
6: **Client side:**
7: **for** client $i \in [N]$ in parallel **do**
8: Set $\mathbf{H}_i^{t,0} = \mathbf{H}_i^{t-1}$ and $\mathbf{W}_i^{t,0} = \mathbf{W}^t$.
9: **for** $r = 1$ **to** Q_1 **do**
10: Update $\mathbf{H}_i^{t,r}$ by (17), and set $\mathbf{W}_i^{t,r} = \mathbf{W}_i^{t,r-1}$.
11: **end for**
12: Compute $Q_2^t = \lfloor\frac{\widehat{Q}}{t}\rfloor + 1$.
13: **for** $r = Q_1 + 1$ **to** $Q^t = Q_1 + Q_2^t$ **do**
14: Update $\mathbf{W}_i^{t,r}$ by (18), and set $\mathbf{H}_i^{t,r} = \mathbf{H}_i^{t,r-1}$.
15: **end for**
16: **end for**
17: Set $\mathbf{W}_i^t = \mathbf{W}_i^{t,Q^t}$ and $\mathbf{H}_i^t = \mathbf{H}_i^{t,Q^t}$.
18: **for** client $i \in \mathcal{S}^t$ in parallel **do**
19: Compute $\widetilde{\mathbf{W}}_i^t$ by (19).
20: Upload $\widetilde{\mathbf{W}}_i^t$ to the PS for next round of aggregation.
21: **end for**
22: **end for**

V. Theoretical Analysis

A. Assumptions

We need the following assumptions to analyze the privacy guarantee and convergence performance of the proposed algorithm.

Assumption 1: Each local cost function F_i is continuously differentiable in both $\mathbf{W}$ and $\mathbf{H}_i$. That is, $\nabla_{H_i}F_i(\mathbf{W}^t, \cdot)$ is Lipschitz continuous with constant $L_{H_i}^t$, and $\nabla_W F_i(\cdot, \mathbf{H}_i^t)$ is Lipschitz continuous with constant $L_{W_i}^t$, i.e., for any $\mathbf{X}, \mathbf{Y}$

$$\|\nabla_{H_i}F_i(\mathbf{W}^t, \mathbf{X}) - \nabla_{H_i}F_i(\mathbf{W}^t, \mathbf{Y})\|_F \le L_{H_i}^t\|\mathbf{X} - \mathbf{Y}\|_F \tag{21a}$$

$$\|\nabla_W F_i(\mathbf{X}, \mathbf{H}_i^t) - \nabla_W F_i(\mathbf{Y}, \mathbf{H}_i^t)\|_F \le L_{W_i}^t\|\mathbf{X} - \mathbf{Y}\|_F. \tag{21b}$$

According to Assumption 1 and [35], $\nabla_W F(\cdot, \mathbf{H}^t)$ is Lipschitz continuous with a constant $L_W^t = (\sum_{i=1}^N (L_{W_i}^t)^2/N)^{1/2}$, together with upper and lower bounds for $L_{H_i}^t$ and $L_{W_i}^t$, i.e.,

$$\overline{L}_W \ge L_{W_i}^t \ge \underline{L}_W > 0,\ \overline{L}_H \ge L_{H_i}^t \ge \underline{L}_H > 0 \quad \forall i, t. \tag{22}$$

Assumption 2: All the local cost functions F_i and their gradients are bounded, i.e., for any $i \in [N]$ and t

$$\|\nabla_W F_i(\mathbf{W}, \mathbf{H}_i; \mathcal{B}_i)\|_F^2 \le G^2 \quad \forall \mathbf{W}, \mathbf{H}_i \ge 0 \tag{23}$$

$$F_i(\mathbf{W}, \mathbf{H}_i) \ge \underline{F} > -\infty \quad \forall \mathbf{W}, \mathbf{H}_i \ge 0 \tag{24}$$

where G is a constant, and $\mathcal{B}_i \subseteq \mathcal{D}_i$ denotes the minibatch dataset.

Assumption 3: For any minibatch dataset $\mathcal{B}_i^t$ with size b that are randomly sampled from dataset $\mathcal{D}_i$, the following equations hold:

$$\mathbb{E}\left[\nabla_W F_i(\mathbf{W}_i^t, \mathbf{H}_i^t; \mathcal{B}_i^t)\right] = \nabla_W F_i(\mathbf{W}_i^t, \mathbf{H}_i^t) \tag{25}$$

$$\mathbb{E}\left[\|\nabla_W F_i(\mathbf{W}_i^t, \mathbf{H}_i^t) - \nabla_W F_i(\mathbf{W}_i^t, \mathbf{H}_i^t; \mathcal{B}_i^t)\|_F^2\right] \le \frac{\phi^2}{b} \tag{26}$$

for any $i \in [N]$ and t, where ϕ is a constant.

Assumption 4: (ζ-non-i.i.d. data) All the local cost functions F_i (cf. (14)) are ζ-non-i.i.d., namely, the following condition holds:

$$\|\nabla_W F_i(\mathbf{W}, \mathbf{H}_i) - \nabla_W F(\mathbf{W}, \mathbf{H})\|_F^2 \le \zeta^2 \quad \forall \mathbf{W}, \mathbf{H} \ge 0 \tag{27}$$

where $\zeta \ge 0$ is a constant.

By following similar spirits to those in [35] and [52], an upper bound ζ is enforced on all the gradients of F_i and F due to the heterogeneity of local data distributions among clients. This bound actually reflects the data's non-i.i.d. degree, which has been extensively utilized in the FL community, particularly for handling nonconvex FL problems.

B. Privacy Analysis

1) Privacy Guarantee: The ℓ_2-norm sensitivity [11] of $\mathbf{W}_i^t$ is stated in the following lemma.

Lemma 2: For any $t \in [R]$ and $i \in [N]$, the ℓ_2-norm sensitivity of uploaded local model $\mathbf{W}_i^t$ is given by

$$s_i^t = \frac{2GQ_2^t}{\eta^t}. \tag{28}$$

Proof: See Appendix B. ■

According to Lemmas 1 and 2, we further come up with the following theorem, which can serve as a guideline for determining the variance of DP noise necessary to fulfill the associated DP-based FL.

Theorem 2: For any client $i \in [N]$, suppose that $\epsilon \leq 1$, $\delta \leq 1$, and the data sampling ratio $q_{i,t} = Q_2^t b/n_i$ (cf. Remark 1). Each entry of $\boldsymbol{\xi}_i^t$ generated follows the Gaussian distribution with zero mean and variance $\sigma_{i,t}^2$, where

$$\sigma_{i,t}^2 = \frac{32G^2\left(Q_2^t\right)^2 q_{i,t}^2 \ln\left(1.25q_{i,t}/\delta\right)}{(\eta^t)^2\epsilon^2}. \tag{29}$$

Then each communication round of the proposed algorithm guarantees (ϵ, δ)-DP.

Proof: In each communication round of the proposed algorithm, each client i performs Q_2^t steps of SGD w.r.t. $\mathbf{W}$ by (18), where the minibatch dataset with size b used is randomly sampled without replacement from local dataset $\mathcal{D}_i$. According to Lemma 1 and Theorem 1, the Gaussian noise with variance

$$\sigma_{i,t}^2 = \frac{2s_{i,t}^2 \ln(1.25/\delta)}{\epsilon^2} \tag{30}$$

can achieve at least $(2q_{i,t}\epsilon, q_{i,t}\delta)$-DP for client i, where $q_{i,t} = Q_2^t b/n_i$ is data sampling ratio for client i. Then, by plugging $s_{i,t}^2$ given by (28) into (30), we obtain

$$\sigma_{i,t}^2 = \frac{4G^2\left(Q_2^t\right)^2 \ln(1.25/\delta)}{(\eta^t)^2\epsilon^2} \quad \forall i \in [N]. \tag{31}$$

By (31), one can achieve an (ϵ, δ)-DP for $\mathbf{W}_i^t$, by replacing ϵ and δ in (31) with $\epsilon/2q_{i,t}$ and $\delta/q_{i,t}$, respectively, thereby leading to (29). ∎

2) Total Privacy Loss: As done in [3], we also use the moments accountant method to estimate the total privacy loss when the algorithm runs R communication rounds.

Theorem 3: Suppose that the client i is uniformly sampled by the PS with a probability p_i and the data sampling ratio $q_{i,t} = Q_2^t b/n_i$ (cf. Remark 1), where $Q_2^t = \lfloor(\widehat{Q}/t)\rfloor + 1$. Then, with noise variance $\sigma_{i,t}^2$ (stated in Theorem 2) used for the generation of the DP noise under (ϵ, δ)-DP at each communication round, an achievable total privacy loss $\bar{\epsilon}_i$ for client i after R communication rounds is given by

$$\bar{\epsilon}_i = c_0 q_{i,t}^2 \epsilon \sqrt{\frac{p_i R}{1 - q_{i,t}}} \quad \forall i \in [N] \tag{32}$$

where c_0 is a constant. ∎

Proof: The proof basically follows that of Theorem 1 reported in [3]. However, we further consider privacy amplification. Thus, the desired result (32) can be obtained by replacing the ϵ with $2q_{i,t}\epsilon$ in the corresponding $\bar{\epsilon}_i$ in [3, Th. 1]. ∎

Theorem 3 shows that the achievable lower bound of total privacy loss $\bar{\epsilon}_i$ for the proposed DP-FedC is tighter than that of the latest reported in [3] and [40] when p and q are appropriately chosen.

Remark 2: When clients are uniformly sampled with a probability p_i, by (32) in Theorem 3, one can infer that Algorithm 1 guarantees $(\mathcal{O}(q\epsilon\sqrt{pR}), \delta)$-DP under R communication rounds, where p and q are given by

$$q = \max_{i,t} \frac{q_{i,t}^2}{\sqrt{1 - q_{i,t}}} \quad \forall i \in [N], t \in [R] \tag{33}$$

$$p = \max_i p_i \quad \forall i \in [N] \tag{34}$$

where $q_{i,t} = Q_2^t b/n_i$.

C. Convergence Analysis

To find some convergence conditions, let us define the following sequence:

$$\overline{\mathbf{W}}^{t,r} = \begin{cases} \frac{1}{K}\sum_{i\in\mathcal{S}^t} \mathbf{W}_i^{t,r}, & \text{when } r \in \left[Q^t - 1\right] \\ \frac{1}{K}\sum_{i\in\mathcal{S}^t} \left(\mathbf{W}_i^{t,Q^t} + \boldsymbol{\xi}_i^t\right), & \text{when } r = Q^t \end{cases} \tag{35}$$

which is actually the instantaneous weighted average of local models. Motivated by [35], let

$$G_H\left(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\right) \triangleq \sum_{i=1}^N (\gamma_i^t)^2 \Big\|\mathbf{H}_i^{t,r} - \Big[\mathbf{H}_i^{t,r} - \frac{1}{\gamma_i^t}\nabla_{H_i} F_i\left(\overline{\mathbf{W}}^{t,r}, \mathbf{H}_i^{t,r}\right)\Big]^+\Big\|_F^2 \quad \forall r \in [Q_1] \tag{36}$$

$$G_W\left(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\right) \triangleq \|\nabla_W F\left(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\right)\|_F^2 \quad \forall r \in \left[Q^t\right] \setminus [Q_1]. \tag{37}$$

If $G_H(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}) = 0$ and $G_W(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}) = 0$, then $(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r})$ is a stationary-point solution of problem (13). The main theoretical result for the DP-FedC is the following theorem.

Theorem 4: Let R be the total number of communication rounds and $T = RQ_1 + \sum_{t=1}^R Q_2^t$ be the total number of gradient evaluations per client. Moreover, let $Q_2^t = \lfloor(\widehat{Q}/t)\rfloor + 1$, $\gamma_i^t = \alpha_1 L_H^t/2$ and $\eta^t = \alpha_2 L_W^t$, where $\alpha_1 > 1$ and $\alpha_2 \geq Q_2^t(3(1 + \overline{L}_W^2/\underline{L}_W^2))^{1/2}$. Then, under Assumptions 1–4, the sequence $\{(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r})\}$ yielded by Algorithm 1 satisfies

$$\begin{aligned} &\frac{1}{T}\Bigg[\sum_{t=1}^R\sum_{r=1}^{Q_1}\mathbb{E}\Big[G_H\left(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\right)\Big] + \sum_{t=1}^R\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[G_W\left(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\right)\Big]\Bigg] \\ &\leq \frac{2\left(\alpha_1^2\overline{L}_H^2 + 1\right)}{T}\Bigg(\alpha_2\overline{L}_W\left(F\left(\overline{\mathbf{W}}^{1,0}, \mathbf{H}^{1,0}\right) - \underline{F}\right) + \frac{16mkG^2\ln(1.25/\delta)\sum_{t=1}^R\left(Q_2^t\right)^3}{\alpha_2\epsilon^2} + \frac{\overline{L}_W\phi^2\sum_{t=1}^R Q_2^t}{2\alpha_2 Kb\underline{L}_W} \\ &\quad + \zeta^2\left(\frac{\sum_{t=1}^R Q_2^t}{K} + \frac{4N\sum_{t=1}^R C_1^t}{\alpha_2^2 K^2}\right)\Bigg) \end{aligned} \tag{38}$$

where

$$C_1^t = Q_2^t\left(Q_2^t - 1\right)\left(2Q_2^t - 1\right). \tag{39}$$

Proof: See Appendix C. ∎

Theorem 4 provides an upper bound of the average total local SGDs over R communication rounds; the smaller its value, the higher convergence rate and the smaller of the cost function in (13) achieved by Algorithm 1. Based on Theorem 4, we have the following remarks.

Remark 3 (Convergence Rate Analysis): Since $Q_2^t = \lfloor(\widehat{Q}/t)\rfloor + 1$, we have $\sum_{t=1}^R C_1^t$, $\sum_{t=1}^R Q_2^t$ and $\sum_{t=1}^R (Q_2^t)^3$ all in $\mathcal{O}(R)$. According to (38), by setting $Q_1 = \mathcal{O}(\sqrt{R})$, the proposed algorithm converges at a rate of $\mathcal{O}(1/\sqrt{R})$. Furthermore, substituting $T = RQ_1 + \sum_{t=1}^R Q_2^t$ into the bound on the right-hand side of (38), one can infer that the bound decreases with Q_1 rather than Q_2 due to $Q_2^t \to 1$ as t increases, implying faster convergence rate for larger Q_1 on the one hand, and the required DP noise variance $\sigma_{i,t}^2$ given by (29) is insensitive to Q_2^t on the other hand.

Remark 4 (Impact of DP): The larger value of ϵ (or $\bar{\epsilon}$), the smaller the upper bound in (38), implying that the better learning performance (convergence rate and the loss function F) and the weaker required privacy protection level, namely, a privacy–utility tradeoff.

Remark 5 (Impact of Non-i.i.d. Data and PCP): The smaller the value of ζ or the larger the value of K, the smaller the upper bound in (38), implying the smaller degree of non-i.i.d. data or the more clients in PCP, and the better learning performance (faster convergence rate and smaller loss function F).

Remark 6 (Complexity Comparison With Existing FedC Methods): Suppose that all clients participate the model training (N clients), the complexity of federated k-means (FKM) for each local iteration at the client side is $\mathcal{O}(mnkN + mnN(\log N)^2)$, and the complexity at the PS side is $\mathcal{O}(mnk)$ [5]. It can be verified that the per-iteration complexity for the proposed DP-FedC algorithm is $\mathcal{O}((mN+n)k^2 + mnk))$ at the client side, and a complexity order of $\mathcal{O}(mkN)$ at the PS side (shown in Appendix E). As a result, the complexity of the proposed DP-FedC algorithm is smaller than that of FKM since $k < N \ll n$ is true in general. However, the DP-FedC and the FZKM [13] have comparable complexity at both client side and the PS side, simply because they have similar computing procedure, in spite of no complexity analysis reported in [13].

VI. Experiment Results

In this section, in terms of the cost function (i.e., the objective value) in (13) and clustering accuracy, some experimental results are presented to evaluate the performance of the proposed DP-FedC algorithm (Algorithm 1) including comparison with some state-of-the-art FedC algorithms. The experiment is performed using two real datasets and each obtained result is the average over five independent runs with the same randomly generated initial feasible points for all the algorithms under test.

A. Experiment Setup

Datasets: The two real datasets used in the experiment are TCGA [38] and MNIST datasets. Specifically, TCGA dataset was obtained from the Cancer Genome Atlas database which contains the gene expression data of 5314 cancer samples belonging to 20 cancer types. Each data sample in TCGA dataset is a real column vector containing the top-ranked 5000 features selected through Pearson's Chi-Squares Test [35]. The MNIST database contains 60 000 training images of ten handwritten digits and 10 000 test ones. We randomly select 10 000 images from the 60 000 training images as the dataset in our experiment, where each data sample is a real column vector containing 784 features. These two datasets are representative, i.e., one (the other) with large (small) data size but small (large) feature size, also implying challenging unsupervised clustering for both datasets in our experiment.

In the experiment, we distribute the samples of each dataset to $N = 100$ clients in the following two ways.

1) *IID Case:* We follow the data partition method in [30] to obtain balanced and i.i.d. distributed data for the two datasets. To be specific, the i.i.d. distributed data are generated by randomly assigning the data samples to all clients.
2) *Non-IID Case:* For the TCGA dataset, we apply the k-means algorithm to cluster the dataset into 100 clusters, and the data samples belonging to the same cluster is assigned to one client. For the MNIST dataset, we follow the partition method in [26] to obtain distributed data such that each client's dataset only contains two digits, thus yielding a highly unbalanced and non-i.i.d. dataset.

Parameter Setting: In problem (13), if not mentioned specifically, we set the parameters as follows: $k = 10$ for MNIST and $k = 20$ for TCGA, $\mu_w = 0$, $\rho = 10^{-7} \times (\|\mathbf{X}\|_F^2/N)$, and $\mu_h = 10^{-10} \times (\|\mathbf{X}\|_F^2/N)$. As for the parameters in Algorithm 1, the step size $\gamma_i^t = (1/2)L_{H_i}^t$ where $L_{H_i}^t$ is estimated as $\lambda_{\max}((\mathbf{W}_i^{t,0})^\top \mathbf{W}_i^{t,0})$. Analogously, the step size $\eta^t = 5L_W^t$ where L_W^t is estimated as $\lambda_{\max}(\mathbf{H}^{t,Q_1}(\mathbf{H}^{t,Q_1})^\top)$. In all experiments, we assume all clients have the same privacy protection level (i.e., $\epsilon_i = \epsilon$, for all i) and the same total privacy loss budget (i.e., $\bar{\epsilon}_i = \bar{\epsilon}$ for all i). Then, given the total privacy loss $\bar{\epsilon}$, the privacy protection level ϵ at each communication round is obtained by Theorem 3 for $R = 100$ and $\delta = 10^{-4}$. The minibatch dataset size b is set to 50. Other parameters are empirically chosen to our best. All the algorithms under test run until $R = 100$ is reached. Then, the clustering accuracy is calculated as the ratio of the number of correct classifications (no. of columns of all the estimated $\mathbf{H}_i$, $i \in [N]$, i.e., their maximum column entries falling in the correct cluster) to the total number of data (i.e., n).

B. Impact of DP

Fig. 3 depicts the objective value for simplicity [i.e., the value of $F(\mathbf{W}, \mathbf{H})$ in (13)] and the clustering accuracy versus communication round with different values of $\bar{\epsilon}$ for both *IID case* and *non-IID case*, where $K = 30$, $Q_1 = 10$, and $Q_2^t = \lfloor(10/t)\rfloor + 1$. Some observations from Fig. 3(a)–(d), are as follows.

1) The larger the value of $\bar{\epsilon}$ where the results without DP conceptually corresponds to $\bar{\epsilon} \to \infty$, the smaller the objective value and the higher the clustering accuracy

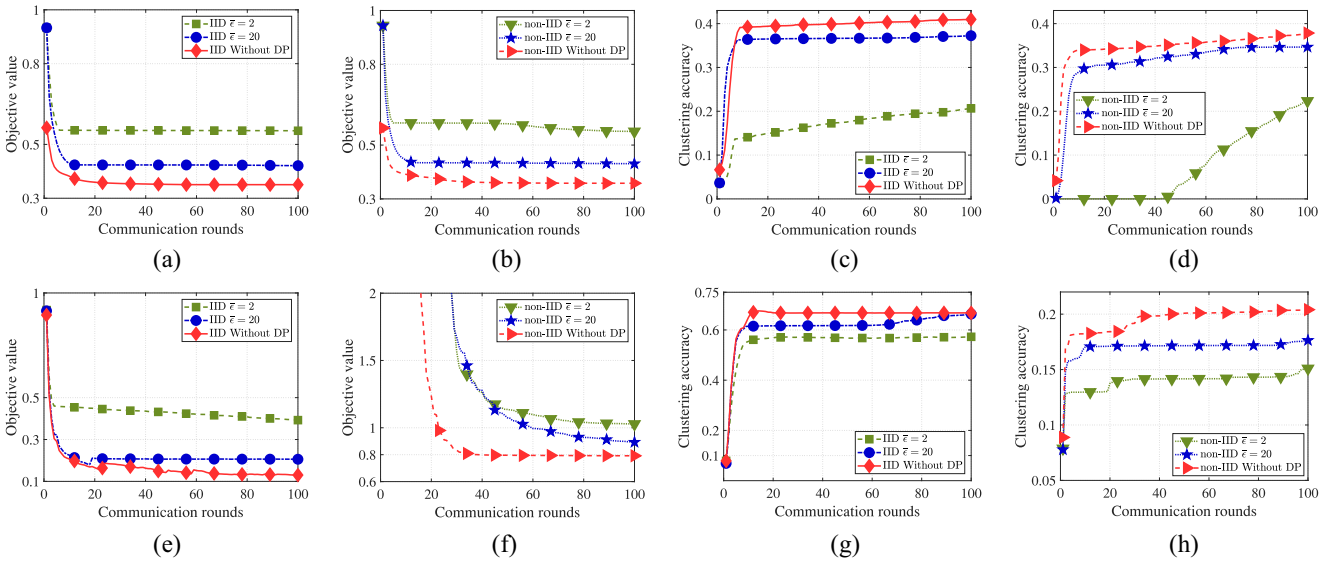


Fig. 3. Objective value and clustering accuracy versus communication rounds of the proposed DP-FedC algorithm for *IID* case and *non-IID* case, where (a)–(d), and (e) and (f), are obtained using the MNIST dataset and TCGA dataset, respectively, for the cases of without DP, and $\bar{\epsilon} \in \{2, 20\}$.

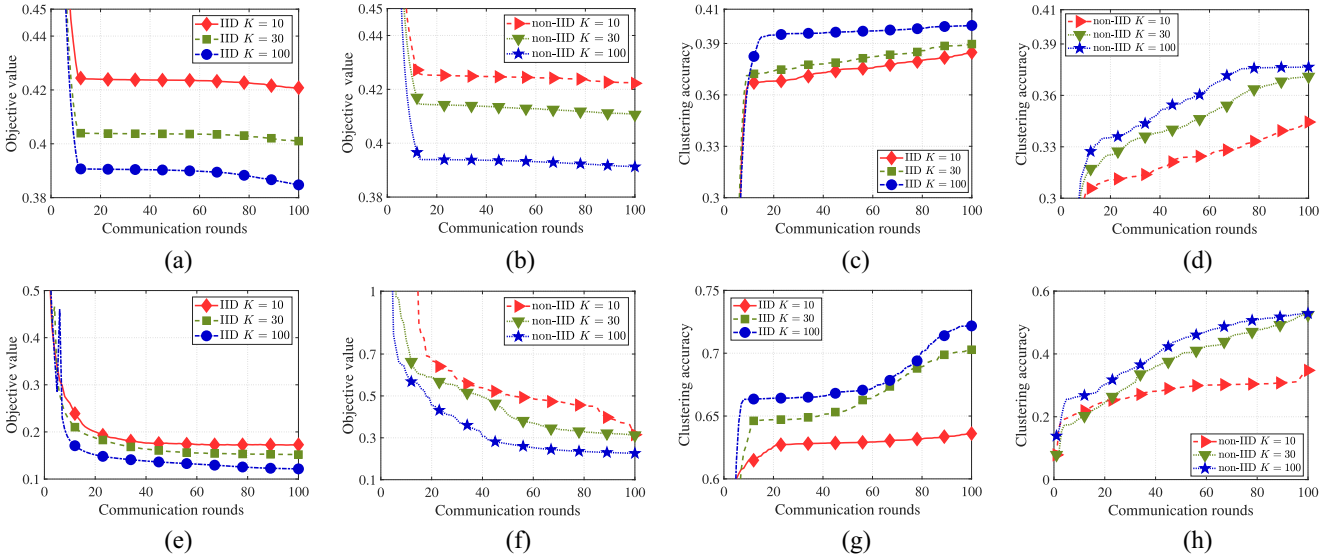


Fig. 4. Objective value and clustering accuracy versus communication rounds of the proposed DP-FedC algorithm for the *IID* case and *non-IID* case, where (a)–(d), and (e) and (f), are obtained using the MNIST dataset and TCGA dataset, respectively, for $\bar{\epsilon} = 20$ and $K \in \{10, 30, 100\}$.

and convergence rate for both *IID case* and *non-IID case*.

2) The objective value is smaller and the clustering accuracy is higher for the *IID case* than for *non-IID case*, and the performance gap between the two cases seems more appreciable in clustering accuracy.

The above two observations also apply to Fig. 3(e)–(h). Moreover, the impact of non-i.i.d. data is more serious on the TCGA dataset. These results are consistent with Remarks 4 and 5, so a proper choice of $\bar{\epsilon}$ value is needed to achieve a good privacy–utility tradeoff.

C. Impact of the Number of Participated Clients (K)

Fig. 4 depicts the convergence performance of DP-FedC versus communication rounds under different values of K with $\bar{\epsilon} = 20$, $Q_1 = 10$, and $Q_2^t = \lfloor (10/t) \rfloor + 1$. It can be seen from Fig. 4(a), (b), (e), and (f), that the objective value is smaller together with faster convergence rate either for larger K or for the *IID* case. This is also true for the clustering accuracy, though the convergence rate on TCGA for the *IID* case is only slightly better than for the *non-IID* case. These results are also consistent with Remark 5.

D. Comparison With Existing Distributed Clustering Methods

We here compare the proposed DP-FedC algorithm with four benchmark algorithms in terms of clustering performance. These algorithms include FKM [5], federated fuzzy k-means (FZKM) [13], distributed k-means++ (DK++) [27], and distributed k-median (DKM) [30]. The first two are state-of-the-art FedC algorithms while the latter two are traditional distributed clustering methods. As mentioned previously, they

TABLE I
PERFORMANCE COMPARISON OF FIVE ALGORITHMS IN TERMS OF CLUSTERING ACCURACY (%)

Method \ Dataset	TCGA (without DP)	MNIST (without DP)	TCGA ($\bar{\epsilon} = 20$)	MNIST ($\bar{\epsilon} = 20$)
DK++ [27]	65.4	42.6	50.1	26.8
DKM [30]	38.7	43.3	31.0	26.1
FKM [5]	70.2	43.2	58.4	31.8
FZKM [13]	72.8	47.1	66.9	36.4
DP-FedC	**76.7**	**50.5**	**72.2**	**43.1**

were basically developed by extending the k-means algorithm and its variants. We add the artificial noise to DP noise that guarantees the (ϵ, δ)-DP at each communication round in the implementation of the above four existing algorithms in our experiment. Then, we apply the proposed algorithm to process the given dataset with parameters $K = 30$, $Q_1 = 10$, $Q_2^t = 5$, and $\bar{\epsilon} = 20$ under the i.i.d. data case. However, the parameters used for the other four algorithms are taken from the associated references together with $K = 30$ and $\bar{\epsilon} = 20$.

The obtained experimental results (for the clustering accuracy) are listed in Table I. It can be seen from this table that the clustering accuracy performances of all the algorithms under test for the case of without DP noise are better than with DP noise used. The performance gap between the two cases for our DP-FedC algorithm is much smaller than for the other algorithms, implying that the proposed algorithm is more robust again DP noise thanks to the privacy amplification strategy applied.

VII. CONCLUSION

We have presented a novel FedC algorithm called DP-FedC (Algorithm 1), which is based on the traditional clustering algorithm k-means and operates according to the computation-aggregation protocol. Specifically, the proposed DP-FedC employs DP-based privacy protection, along with the policies of PCP and multiple local SGD updating steps implemented in the algorithm design. Various characteristics and insights of Algorithm 1 were discovered through theoretical analyses, including the impact of system parameters on privacy amplification, convergence rate, and the impact of data heterogeneity (e.g., non-i.i.d. data) on learning performance. These analytical results can serve as valuable guidelines for practical FL algorithm design, especially when considering the preferred tradeoff between learning performance and the required level of privacy protection. Finally, we provided experimental results on two real datasets to demonstrate the efficacy of the proposed method, along with its superior performance over state-of-the-art FedC algorithms, and its consistency with all the presented analytical results.

APPENDIX A
PROOF OF THEOREM 1

The proof mainly follows the work [23] by considering both data sampling with replacement case and that without replacement case.

Suppose that data subsampling mechanism yields (ϵ', δ')-DP, when $\epsilon \leq 1$ and data are uniformly sampled with replacement, the data subsampling mechanism guarantees $(\ln(1 + q(\exp(\epsilon) - 1), q\delta)$-DP [23], then, we have $\delta' = q\delta$ and

$$\begin{aligned}\epsilon' &= \ln(1 + q(\exp(\epsilon) - 1)) \\ &\overset{(a)}{\leq} q(\exp(\epsilon) - 1) \overset{(b)}{\leq} 2q\epsilon \end{aligned} \tag{40}$$

where (a) and (b) hold because $\ln(1+x) \leq x$ and $\exp(x)-1 \leq 2x$ when $0 < x \leq 1$ [53].

When data are uniformly sampled without replacement, we still have $\delta' = q\delta$ and ϵ' becomes [23]

$$\begin{aligned}\epsilon' &= \ln\left(1 + \left(1 - \left(1 - \frac{1}{n}\right)\right)^b (\exp(\epsilon) - 1)\right) \\ &\overset{(a)}{\leq} \ln\left(1 + q(\exp(\epsilon) - 1)\right) \overset{(b)}{\leq} 2q\epsilon \end{aligned} \tag{41}$$

where (b) follows because of (40), and (a) holds since:

$$\left(1 - \left(1 - \tfrac{1}{n}\right)\right)^b \leq \tfrac{b}{n} = q. \tag{42}$$

By combining (41) and (42), we obtain $\epsilon' \leq 2q\epsilon$ for data sampling without replacement.

Then, when $q \geq 1/2$ (i.e., $2q\epsilon > \epsilon$), there is no privacy amplification. In this case, we have

$$\epsilon' = \epsilon. \tag{43}$$

Therefore, by combining (40), (41), and (43), we have $\epsilon' = \min(2q\epsilon, \epsilon)$. Thus, we complete the proof. ■

APPENDIX B
PROOF OF LEMMA 2

Assume $\mathcal{D}_i$ and $\mathcal{D}_i'$ are the neighboring datasets that differ in only one data sample. Without loss of generality, let u_i be the unique different element between $\mathcal{D}_i$ and $\mathcal{D}_i'$, i.e., $\mathcal{D}_i' \cup \{u_i\} = \mathcal{D}_i \cup \{u_i\}$. For clarity of the following proof, let us make the following notational correspondences: $\mathbf{W}_i^{t,r} \leftrightarrow \mathbf{W}_{\mathcal{D}_i}^{t,r}$, $\mathbf{H}_i^{t,r} \leftrightarrow \mathbf{H}_{\mathcal{D}_i}^{t,r}$, and $\mathcal{B}_i^{t,r} \leftrightarrow \mathcal{B}_{\mathcal{D}_i}^{t,r}$. Then, for any $r \in [Q^t] \setminus [Q_1]$, the ℓ_2-sensitivity [11] of $\mathbf{W}_i^t$ is calculated by

$$\begin{aligned} s_i^t &= \max_{\mathcal{D}_i, \mathcal{D}_i'} \left\| \mathbf{W}_{\mathcal{D}_i}^t - \mathbf{W}_{\mathcal{D}_i'}^t \right\| \\ &= \max_{\mathcal{D}_i, \mathcal{D}_i'} \left\| \sum_{r=Q_1+1}^{Q^t} \mathbf{W}_{\mathcal{D}_i}^{t,r-1} - \frac{\nabla_W F_i\left(\mathbf{W}_{\mathcal{D}_i}^{t,r-1}, \mathbf{H}_{\mathcal{D}_i}^{t,r-1}; \mathcal{B}_{\mathcal{D}_i}^{t,r}\right)}{\eta^t} \right. \\ &\quad \left. - \sum_{r=Q_1+1}^{Q^t} \left(\mathbf{W}_{\mathcal{D}_i'}^{t,r-1} - \frac{\nabla_W F_i\left(\mathbf{W}_{\mathcal{D}_i'}^{t,r-1}, \mathbf{H}_{\mathcal{D}_i'}^{t,r-1}; \mathcal{B}_{\mathcal{D}_i'}^{t,r}\right)}{\eta^t} \right) \right\| \end{aligned}$$

$$= \max_{\mathcal{D}_i,\mathcal{D}_i'} \Bigg\| \Bigg(\mathbf{W}_{\mathcal{D}_i}^{t,Q_1} - \frac{\nabla_W F_i\Big(\mathbf{W}_{\mathcal{D}_i}^{t,Q_1}, \mathbf{H}_{\mathcal{D}_i}^{t,Q_1}; \mathcal{B}_{\mathcal{D}_i}^{t,Q_1+1}\Big)}{\eta^t} - \cdots - \frac{\nabla_W F_i\Big(\mathbf{W}_{\mathcal{D}_i}^{t,Q^t-1}, \mathbf{H}_{\mathcal{D}_i}^{t,Q^t-1}; \mathcal{B}_{\mathcal{D}_i}^{t,Q^t}\Big)}{\eta^t} \Bigg) - \Bigg(\mathbf{W}_{\mathcal{D}_i'}^{t,Q_1} - \frac{\nabla_W F_i\Big(\mathbf{W}_{\mathcal{D}_i'}^{t,Q_1}, \mathbf{H}_{\mathcal{D}_i'}^{t,Q_1}; \mathcal{B}_{\mathcal{D}_i'}^{t,Q_1+1}\Big)}{\eta^t} - \cdots - \frac{\nabla_W F_i\Big(\mathbf{W}_{\mathcal{D}_i'}^{t,Q^t-1}, \mathbf{H}_{\mathcal{D}_i'}^{t,Q^t-1}; \mathcal{B}_{\mathcal{D}_i'}^{t,Q^t}\Big)}{\eta^t} \Bigg) \Bigg\|$$

$$\overset{(a)}{\leq} \frac{2GQ_2^t}{\eta^t} \tag{44}$$

where (a) holds because of Assumption 2, and $\mathbf{W}_{\mathcal{D}_i}^{t,Q_1} = \mathbf{W}_{\mathcal{D}_i'}^{t,Q_1}$ always holds. ∎

Appendix C
Proof of Theorem 4

According to (35) and (18), we have

$$\overline{\mathbf{W}}^{t,r} = \overline{\mathbf{W}}^{t,r-1} - \tfrac{1}{K\eta^t} \textstyle\sum_{i\in\mathcal{S}^t} \nabla_W F_i\Big(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^{t,r-1}; \mathcal{B}_i^{t,r}\Big). \tag{45}$$

Objective Descent w.r.t. $\boldsymbol{H}$: According to [54, Lemma 3.2] and setting $\gamma_i^t = \alpha_1 L_H^t/2 \leq \alpha_1 \overline{L}_H/2$ where $\alpha_1 > 1$, we have

$$F_i\Big(\overline{\mathbf{W}}^{t,r}, \mathbf{H}_i^{t,r}\Big) - F_i\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}_i^{t,r-1}\Big) \leq -\frac{\alpha_1 - 1}{2}\overline{L}_H \|\mathbf{H}_i^{t,r-1} - \mathbf{H}_i^{t,r}\|_F^2 \quad \forall r \in [Q_1]. \tag{46}$$

Taking expectation over two sides of (46) and then summing up from $r = 1$ to Q_1 yields

$$\mathbb{E}\Big[F_i\Big(\overline{\mathbf{W}}^{t,Q_1}, \mathbf{H}_i^{t,Q_1}\Big)\Big] - \mathbb{E}\Big[F_i\Big(\overline{\mathbf{W}}^{t,0}, \mathbf{H}_i^{t,0}\Big)\Big] \leq -\frac{\alpha_1 - 1}{2}\overline{L}_H \sum_{r=1}^{Q_1} \mathbb{E}\big[\|\mathbf{H}_i^{t,r-1} - \mathbf{H}_i^{t,r}\|_F^2\big] \quad \forall r \in [Q_1]. \tag{47}$$

By taking the summation over two sides of (47) from $i = 1$ to N, the objective function F descends with local updates of $\mathbf{H}$ is given by

$$\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q_1}, \mathbf{H}^{t,Q_1}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,0}, \mathbf{H}^{t,0}\Big)\Big] \leq -\frac{\alpha_1 - 1}{2}\overline{L}_H \sum_{r=1}^{Q_1}\sum_{i=1}^{N} \mathbb{E}\big[\|\mathbf{H}_i^{t,r-1} - \mathbf{H}_i^{t,r}\|_F^2\big] \quad \forall r \in [Q_1]. \tag{48}$$

Objective Descent w.r.t. $\boldsymbol{W}$: Since $\mathbf{H}_i^{t,r} = \mathbf{H}_i^{t,r-1}$ (cf. line 14 in Algorithm 1) and $\nabla_W F(\cdot, \mathbf{H}^{t,Q})$ is Lipschitz continuous under Assumption 1. Then, by the descent lemma [54, Lemma 3.1], when $r \in [Q^t - 1] \setminus [Q_1]$, we have

$$\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\Big)\Big] \leq \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big] + \frac{L_W^t}{2}\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\|_F^2\Big] + \mathbb{E}\Big[\Big\langle \nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big), \overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\Big\rangle\Big]. \tag{49}$$

When $r = Q^t$, by Algorithm 1, (49) becomes

$$\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t}, \mathbf{H}^{t,Q^t}\Big)\Big] \leq \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t-1}, \mathbf{H}^{t,Q^t-1}\Big)\Big] + \underbrace{\frac{L_W^t}{2}\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,Q^t} - \overline{\mathbf{W}}^{t,Q^t-1} + \boldsymbol{\xi}^t\|_F^2\Big]}_{\triangleq \text{(S.1)}} + \mathbb{E}\Big[\Big\langle \nabla_W F\Big(\overline{\mathbf{W}}^{t,Q^t-1}, \mathbf{H}^{t,Q^t-1}\Big), \overline{\mathbf{W}}^{t,Q^t} - \overline{\mathbf{W}}^{t,Q^t-1}\Big\rangle\Big] \tag{50}$$

where $\boldsymbol{\xi}^t = (1/K)\sum_{i=i}^K \boldsymbol{\xi}_i^t$. The (S.1) can be further bounded by

$$\begin{aligned}\text{(S.1)} &= \frac{L_W^t}{2}\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,Q^t} - \overline{\mathbf{W}}^{t,Q^t-1} + \boldsymbol{\xi}^t\|_F^2\Big] \\ &= \frac{L_W^t}{2}\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,Q^t-1} - \overline{\mathbf{W}}^{t,Q^t}\|_F^2\Big] + \frac{L_W^t}{2}\mathbb{E}\Big[\|\boldsymbol{\xi}^t\|_F^2\Big] \\ &\overset{(a)}{\leq} \frac{L_W^t}{2}\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,Q^t-1} - \overline{\mathbf{W}}^{t,Q^t}\|_F^2\Big] + \frac{16mkG^2\ln(1.25/\delta)\big(Q_2^t\big)^2}{\alpha_2\eta^t\epsilon^2}\end{aligned} \tag{51}$$

where (a) holds from $\eta^t = \alpha_2 L_W^t$ and

$$\mathbb{E}\big[\|\boldsymbol{\xi}^t\|_F^2\big] \overset{(a)}{=} \frac{mk}{K}\sum_{i=1}^K \frac{32G^2\big(Q_2^t\big)^2 q_{i,t}^2 \ln\big(1.25q_{i,t}/\delta\big)}{(\eta^t)^2\epsilon^2} \overset{(b)}{\leq} \frac{32mkG^2\big(Q_2^t\big)^2\ln(1.25/\delta)}{(\eta^t)^2\epsilon^2}. \tag{52}$$

In (52), (a) follows from (29). (b) holds because of $q_{i,t} \leq 1$. By (49)–(51), for $r \in [Q^t] \setminus [Q_1]$, we have

$$\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\Big)\Big] \leq \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big] + \frac{L_W^t}{2}\underbrace{\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\|_F^2\Big]}_{\triangleq \text{(S.2)}} + \underbrace{\mathbb{E}\Big[\Big\langle \nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big), \overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\Big\rangle\Big]}_{\triangleq \text{(S.3)}} + \frac{16mkG^2\big(Q_2^t\big)^2\ln(1.25/\delta)}{\alpha_2\eta^t\epsilon^2}. \tag{53}$$

The terms (S.2) and (S.3) can be bounded by the following Lemma 3 (proved in Appendix D-A) and Lemma 4 (proved in Appendix D-B), respectively.

Lemma 3: For any t and $r \in [Q^t - 1] \setminus [Q_1]$, we have

$$\mathbb{E}\Big[\Big\|\overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\Big\|_F^2\Big] \leq \frac{\phi^2}{Kb(\eta^t)^2} + \frac{1}{(\eta^t)^2}\mathbb{E}\Bigg[\Bigg\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^{t,r-1}\Big)\Bigg\|^2\Bigg]. \tag{54}$$

Lemma 4: For any t and $r \in [Q^t - 1] \setminus [Q_1]$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big), \overline{\mathbf{W}}^{t,r} - \overline{\mathbf{W}}^{t,r-1}\Big\rangle\Big] \\
&= -\frac{1}{2\eta^t}\mathbb{E}\Bigg[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big\|_F^2 \\
&\quad + \Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^{t,r-1}\Big)\Big\|_F^2\Bigg] \\
&\quad + \frac{\zeta^2}{K\eta^t} + \frac{1}{K\eta^t}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big]. \quad (55)
\end{aligned}$$

Thus, substituting (54) and (55) into (53) gives rise to

$$\begin{aligned}
&\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big] \\
&\le -\frac{1}{2\eta^t}\mathbb{E}\Bigg[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big\|_F^2\Bigg] + \frac{L_W^t\phi^2}{2Kb(\eta^t)^2} + \frac{\zeta^2}{K\eta^t} \\
&\quad + \Big(\frac{L_W^t}{2(\eta^t)^2} - \frac{1}{2\eta^t}\Big)\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1}, \mathbf{H}_i^{t,r-1}\Big)\Big\|^2\Bigg] \\
&\quad + \frac{1}{K\eta^t}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big] \\
&\quad + \frac{16mkG^2\big(Q_2^t\big)^2\ln(1.25/\delta)}{\alpha_2\eta^t\epsilon^2} \\
&\overset{(a)}{\le} -\frac{1}{2\eta^t}\mathbb{E}\Bigg[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big\|_F^2\Bigg] + \frac{\overline{L}_W\phi^2}{2Kb(\eta^t)^2} + \frac{\zeta^2}{K\eta^t} \\
&\quad + \frac{1}{K\eta^t}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big] \\
&\quad + \frac{16mkG^2\big(Q_2^t\big)^2\ln(1.25/\delta)}{\alpha_2\eta^t\epsilon^2} \quad (56)
\end{aligned}$$

where (a) follows due to $\eta^t = \alpha_2 L_W^t \ge L_W^t$ and $L_{W_i}^t \le \overline{L}_W$. Then, rearranging the two sides of (56) yields

$$\begin{aligned}
&\mathbb{E}\Big[\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\|_F^2\Big] \\
&\le 2\eta^t\Big(\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,r}, \mathbf{H}^{t,r}\Big)\Big]\Big) \\
&\quad + \frac{2}{K}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big] + \frac{\overline{L}_W\phi^2}{Kb\eta^t} \\
&\quad + \frac{32mkG^2\big(Q_2^t\big)^2\ln(1.25/\delta)}{\alpha_2\epsilon^2} + \frac{2\zeta^2}{K}. \quad (57)
\end{aligned}$$

Summing (57) up from $r = Q_1 + 1$ to Q^t yields

$$\begin{aligned}
&\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\|_F^2\Big] \\
&\le 2\eta^t\Big(\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q_1}, \mathbf{H}^{t,Q_1}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t}, \mathbf{H}^{t,Q^t}\Big)\Big]\Big) \\
&\quad + \frac{2}{K}\underbrace{\sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big]}_{\triangleq(\text{S.4})} \\
&\quad + \frac{32mkG^2\big(Q_2^t\big)^3\ln(1.25/\delta)}{\alpha_2\epsilon^2} + \frac{2\zeta^2 Q_2^t}{K} + \frac{Q_2^t\overline{L}_W\phi^2}{Kb\eta^t}. \quad (58)
\end{aligned}$$

The term (S.4) can be bounded with the following lemma, which is proved in Appendix D-C.

Lemma 5: Let $\alpha_2 \ge Q_2^t\sqrt{3(1 + \overline{L}_W^2/\underline{L}_W^2)}$. For any t and $r \in [Q^t] \setminus [Q_1]$, it holds that

$$\sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}\big(L_{W_i}^t\big)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1} - \mathbf{W}_i^{t,r-1}\|_F^2\Big] \le \frac{4N\zeta^2 C_1^t}{K\alpha_2^2} \quad (59)$$

where $C_1^t \triangleq Q_2^t(Q_2^t - 1)(2Q_2^t - 1)$.

By applying Lemma 5 and plugging (59) into (58), we have

$$\begin{aligned}
&\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\|_F^2\Big] \\
&\le 2\eta^t\Big(\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q_1}, \mathbf{H}^{t,Q_1}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t}, \mathbf{H}^{t,Q^t}\Big)\Big]\Big) \\
&\quad + \frac{32mkG^2\big(Q_2^t\big)^3\ln(1.25/\delta)}{\alpha_2\epsilon^2} + \frac{2\zeta^2 Q_2^t}{K} \\
&\quad + \frac{Q_2^t\overline{L}_W\phi^2}{Kb\eta^t} + \frac{8N\zeta^2 C_1^t}{K^2\alpha_2^2}. \quad (60)
\end{aligned}$$

Combining (48) and (60) yields

$$\begin{aligned}
&\sum_{r=1}^{Q_1}\sum_{i=1}^{N}\mathbb{E}\Big[\|\mathbf{H}_i^{t,r-1} - \mathbf{H}_i^{t,r}\|_F^2\Big] \\
&\quad + \sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\|_F^2\Big] \\
&\overset{(a)}{\le} 2\eta^t\Big(\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,0}, \mathbf{H}^{t,0}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t}, \mathbf{H}^{t,Q^t}\Big)\Big]\Big) \\
&\quad + \frac{32mkG^2\big(Q_2^t\big)^3\ln(1.25/\delta)}{\alpha_2\epsilon^2} + \frac{2\zeta^2 Q_2^t}{K} \\
&\quad + \frac{Q_2^t\overline{L}_W\phi^2}{Kb\eta^t} + \frac{8N\zeta^2 C_1^t}{K^2\alpha_2^2} \quad (61)
\end{aligned}$$

where (a) holds because of $\eta^t \ge 1/((\alpha_1 - 1)\,\overline{L}_H)$.

Derivation of the Main Result: We next derive the convergence in terms of the optimal gap functions in (36) and (37). From (61) and $\gamma_i^t = \alpha_1 L_H^t/2$ and $\eta^t = \alpha_2 L_W^t$, we have

$$\begin{aligned}
&\sum_{r=1}^{Q_1}\mathbb{E}\Big[G_H\Big(\overline{\mathbf{W}}^{t,r-1}, \mathbf{H}^{t,r-1}\Big)\Big] \\
&= \sum_{r=1}^{Q_1}\sum_{i=1}^{N}\big(\gamma_i^t\big)^2\mathbb{E}\Big[\|\mathbf{H}_i^{t,r-1} - \mathbf{H}_i^{t,r}\|_F^2\Big] \\
&\overset{(a)}{\le} 2\alpha_1^2\overline{L}_H\Bigg(\alpha_2\overline{L}_W\mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,0}, \mathbf{H}^{t,0}\Big)\Big] - \mathbb{E}\Big[F\Big(\overline{\mathbf{W}}^{t,Q^t}, \mathbf{H}^{t,Q^t}\Big)\Big] \\
&\quad + \frac{16mkG^2\big(Q_2^t\big)^3\ln(1.25/\delta)}{\alpha_2\epsilon^2} + \frac{\zeta^2 Q_2^t}{K} \\
&\quad + \frac{Q_2^t\overline{L}_W\phi^2}{2Kb\underline{L}_W\alpha_2} + \frac{4N\zeta^2 C_1^t}{K^2\alpha_2^2}\Bigg) \quad (62)
\end{aligned}$$

where (a) follows because $\gamma_i^t \le \alpha_1 \bar{L}_H/2$ and $\alpha_2 \underline{L}_W \le \eta^t \le \alpha_2 \bar{L}_W$. Then, summing (62) up from $t=1$ to R yields

$$\begin{aligned}
&\sum_{t=1}^{R}\sum_{r=1}^{Q_1}\mathbb{E}\Big[G_H\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big]\\
&\le 2\alpha_1^2\bar{L}_H^2\Bigg(\alpha_2\bar{L}_W\big(F\Big(\overline{\mathbf{W}}^{1,0},\mathbf{H}^{1,0}\Big)-\underline{F}\big)\\
&\quad+\frac{16mkG^2\ln(1.25/\delta)\sum_{t=1}^{R}\big(Q_2^t\big)^3}{\alpha_2\epsilon^2}+\frac{\zeta^2\sum_{t=1}^{R}Q_2^t}{K}\\
&\quad+\frac{\bar{L}_W\phi^2\sum_{t=1}^{R}Q_2^t}{2Kb\underline{L}_W\alpha_2}+\frac{4N\zeta^2\sum_{t=1}^{R}C_1^t}{K^2\alpha_2^2}\Bigg).
\end{aligned}\tag{63}$$

Similarly, from (61), we have

$$\begin{aligned}
&\sum_{t=1}^{R}\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Bigg[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big\|_F^2\Bigg]\\
&=\sum_{t=1}^{R}\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[G_W\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big]\Big]\\
&\le 2\alpha_2\bar{L}_W\big(F\Big(\overline{\mathbf{W}}^{1,0},\mathbf{H}^{1,0}\Big)-\underline{F}\big)\\
&\quad+\frac{32mkG^2\ln(1.25/\delta)\sum_{t=1}^{R}\big(Q_2^t\big)^3}{\alpha_2\epsilon^2}+\frac{2\zeta^2\sum_{t=1}^{R}Q_2^t}{K}\\
&\quad+\frac{\bar{L}_W\phi^2\sum_{t=1}^{R}Q_2^t}{Kb\underline{L}_W\alpha_2}+\frac{8N\zeta^2\sum_{t=1}^{R}C_1^t}{K^2\alpha_2^2}.
\end{aligned}\tag{64}$$

By combining (63) and (64), and then dividing two sides of summation result by $T=RQ_1+\sum_{t=1}^{R}Q_2^t$ yields

$$\begin{aligned}
&\frac{1}{T}\Bigg[\sum_{t=1}^{R}\sum_{r=1}^{Q_1}\mathbb{E}\Big[G_H\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big]\\
&\quad+\sum_{t=1}^{R}\sum_{r=Q_1+1}^{Q^t}\mathbb{E}\Big[G_W\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big]\Bigg]\\
&\le\frac{2\Big(\alpha_1^2\bar{L}_H^2+1\Big)}{T}\Bigg(\alpha_2\bar{L}_W\big(F\Big(\overline{\mathbf{W}}^{1,0},\mathbf{H}^{1,0}\Big)-\underline{F}\big)\\
&\quad+\frac{16mkG^2\ln(1.25/\delta)\sum_{t=1}^{R}\big(Q_2^t\big)^3}{\alpha_2\epsilon^2}+\frac{\zeta^2\sum_{t=1}^{R}Q_2^t}{K}\\
&\quad+\frac{\bar{L}_W\phi^2\sum_{t=1}^{R}Q_2^t}{2Kb\underline{L}_W\alpha_2}+\frac{4N\zeta^2\sum_{t=1}^{R}C_1^t}{K^2\alpha_2^2}\Bigg).
\end{aligned}\tag{65}$$

This completes the proof. ■

APPENDIX D
PROOFS OF KEY LEMMAS FOR THEOREM 4

A. Proof of Lemma 3

According to (45), we have

$$\begin{aligned}
&\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r}-\overline{\mathbf{W}}^{t,r-1}\|_F^2\Big]\\
&=\frac{1}{(\eta^t)^2}\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1};\mathcal{B}_i^{t,r}\Big)\Big\|^2\Bigg]\\
&\overset{(a)}{=}\frac{1}{(\eta^t)^2}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\Big(\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1};\mathcal{B}_i^{t,r}\Big)\\
&\quad-\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big)\Big\|^2\Big]\\
&\quad+\frac{1}{(\eta^t)^2}\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|^2\Bigg]\\
&\overset{(b)}{=}\frac{1}{(\eta^t)^2K^2}\mathbb{E}\Big[\sum_{i\in\mathcal{S}^t}\Big\|\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1};\mathcal{B}_i^{t,r}\Big)\\
&\quad-\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|^2\Big]\\
&\quad+\frac{1}{(\eta^t)^2}\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|^2\Bigg]\\
&\overset{(c)}{\le}\frac{1}{(\eta^t)^2}\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|^2\Bigg]+\frac{\phi^2}{Kb(\eta^t)^2}
\end{aligned}\tag{66}$$

where (a) follows because $\mathbb{E}[\|\mathbf{Z}\|^2]=\mathbb{E}[\|\mathbf{Z}-\mathbb{E}[\mathbf{Z}]\|^2]+\|\mathbb{E}[\mathbf{Z}]\|^2$; (b) follows because $\nabla_W F_i(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1};\mathcal{B}_i^{t,r})-\nabla_W F_i(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1})$ is independent across the clients; and (c) holds due to Assumption 3. ■

B. Proof of Lemma 4

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big),\overline{\mathbf{W}}^{t,r}-\overline{\mathbf{W}}^{t,r-1}\Big\rangle\Big]\\
&\overset{(a)}{=}-\frac{1}{\eta^t}\mathbb{E}\Big[\Big\langle\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big),\\
&\quad\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1};\mathcal{B}_i^{t,r}\Big)\Big\rangle\Big]\\
&\overset{(b)}{=}-\frac{1}{\eta^t}\mathbb{E}\Big[\Big\langle\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\\
&\quad\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\rangle\Big]\\
&\overset{(c)}{=}-\frac{1}{2\eta^t}\mathbb{E}\Big[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big\|_F^2\Big]\\
&\quad-\frac{1}{2\eta^t}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|_F^2\Big]\\
&\quad+\frac{1}{2\eta^t}\mathbb{E}\Big[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\\
&\quad-\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|_F^2\Big]
\end{aligned}\tag{67}$$

where (a) holds due to (18); (b) follows from Assumption 3; and (c) follows from the basic identity $\langle\mathbf{Z}_1,\mathbf{Z}_2\rangle=(1/2)(\|\mathbf{Z}_1\|^2+\|\mathbf{Z}_2\|^2-\|\mathbf{Z}_1-\mathbf{Z}_2\|^2)$.

The last term in (67) can be further bounded by

$$\begin{aligned}
&\mathbb{E}\Bigg[\Big\|\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\\
&\quad-\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\Big(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\Big)\Big\|_F^2\Bigg]\\
&=\mathbb{E}\Bigg[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\Big(\nabla_W F\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)-\nabla_W F_i\Big(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\Big)\Big)\Big\|_F^2\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&+\nabla_W F_i\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right)-\nabla_W F_i\left(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\right)\Big)\Big\|_F^2\Big]\\
&\le \frac{2}{K^2}\mathbb{E}\Big[\Big\|\sum_{i\in\mathcal{S}^t}\Big(\nabla_W F\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right)\\
&\quad-\nabla_W F_i\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right)\Big)\Big\|_F^2\Big]\\
&\quad+\frac{2}{K^2}\mathbb{E}\Big[\Big\|\sum_{i\in\mathcal{S}^t}\Big(\nabla_W F_i\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right)\\
&\quad-\nabla_W F_i\left(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\right)\Big)\Big\|_F^2\Big]\\
&\overset{(a)}{\le}\frac{2\zeta^2}{K}+\frac{2}{K}\mathbb{E}\Big[\sum_{i\in\mathcal{S}^t}\left(L_{W_i}^t\right)^2\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\|_F^2\Big]\\
&\le\frac{2\zeta^2}{K}+\frac{2}{K}\sum_{i=1}^{N}\left(L_{W_i}^t\right)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\|_F^2\Big] \qquad (68)
\end{aligned}
$$

where the first term in the RHS of (a) comes from Assumption 4, and the second term in the RHS of (73) follows because of Assumption 1. Then, plugging (68) into (67) yields

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\langle\nabla_W F\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right),\overline{\mathbf{W}}^{t,r}-\overline{\mathbf{W}}^{t,r-1}\Big\rangle\Big]\\
&\le-\frac{1}{2\eta^t}\mathbb{E}\Big[\|\nabla_W F\left(\overline{\mathbf{W}}^{t,r-1},\mathbf{H}^{t,r-1}\right)\|_F^2\Big]\\
&\quad-\frac{1}{2\eta^t}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\left(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,r-1}\right)\Big\|_F^2\Big]\\
&\quad+\frac{\zeta^2}{K\eta^t}+\frac{1}{K\eta^t}\sum_{i=1}^{N}\left(L_{W_i}^t\right)^2\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\|_F^2\Big]. \qquad (69)
\end{aligned}
$$

Thus, we complete the proof. ■

C. Proof of Lemma 5

According to the definition of $\overline{\mathbf{W}}^{t,r-1}$, for $\forall r\in[Q^t]\setminus[Q_1]$, we have

$$
\begin{aligned}
\overline{\mathbf{W}}^{t,r-1}&=\frac{1}{K}\sum_{i\in\mathcal{S}^t}\mathbf{W}_i^{t,r-1}\\
&\overset{(a)}{=}\frac{1}{K}\sum_{i\in\mathcal{S}^t}\Big(\mathbf{W}^t-\frac{1}{\eta^t}\sum_{j=Q_1}^{r-1}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\Big)\\
&=\mathbf{W}^t-\frac{1}{\eta^t K}\sum_{j=Q_1}^{r-1}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right) \qquad (70)
\end{aligned}
$$

where (a) is obtained by applying (18), that is

$$
\mathbf{W}_i^{t,r-1}=\mathbf{W}^t-\frac{1}{\eta^t}\sum_{j=Q_1}^{r-1}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right). \qquad (71)
$$

As a result, by (70) and (71), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\|_F^2\Big]\\
&=\mathbb{E}\Big[\Big\|\mathbf{W}^t-\frac{1}{\eta^t K}\sum_{j=Q_1}^{r-1}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\\
&\quad-\Big(\mathbf{W}^t-\frac{1}{\eta^t}\sum_{j=Q_1}^{r-1}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\Big)\Big\|_F^2\Big]\\
&=\frac{1}{(\eta^t)^2}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{j=Q_1}^{r-1}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\\
&\quad-\sum_{j=Q_1}^{r-1}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\Big\|_F^2\Big]\\
&\le\frac{(r-Q_1)}{(\eta^t)^2}\sum_{j=Q_1}^{r-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j}\right)\\
&\quad-\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1};\mathcal{B}_k^{t,j}\right)\Big\|_F^2\Big]\\
&\overset{(b)}{=}\frac{(r-Q_1)}{(\eta^t)^2}\sum_{j=Q_1}^{r-1}\Big\|\frac{1}{K}\sum_{i\in\mathcal{S}^t}\Big(\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1}\right)\\
&\quad-\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\Big)\Big\|_F^2\\
&\overset{(c)}{\le}\frac{(r-Q_1)}{(\eta^t)^2K}\sum_{j=Q_1}^{r-1}\sum_{i=1}^{N}\Big\|\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1}\right)\\
&\quad-\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\Big\|_F^2 \qquad (72)
\end{aligned}
$$

where (b) holds since $\nabla_W F_i(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1};\mathcal{B}_i^{t,j})-\nabla_W F_k(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1};\mathcal{B}_k^{t,j})$ is independent across the clients; (c) follows by using the inequality $\|\sum_{i=1}^N\mathbf{z}_i\|^2\le N\sum_{i=1}^N\|\mathbf{z}_i\|^2$ for any vectors $\mathbf{z}_i$ and any positive integer N. Then, the term $\|\nabla_W F_i(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1})-\nabla_W F_k(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1})\|_F^2$ in (72) can be bounded by

$$
\begin{aligned}
&\Big\|\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1}\right)-\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\Big\|_F^2\\
&\le\Big\|\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1}\right)-\nabla_W F_i\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_i^{t,j-1}\right)\\
&\quad+\nabla_W F_i\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_i^{t,j-1}\right)-\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)\\
&\quad+\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)-\Big(\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\\
&\quad-\nabla_W F_k\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_k^{t,j-1}\right)+\nabla_W F_k\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\\
&\quad-\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)+\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)\Big)\Big\|_F^2\\
&\le 4\Big\|\nabla_W F_i\left(\mathbf{W}_i^{t,j-1},\mathbf{H}_i^{t,j-1}\right)-\nabla_W F_i\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_i^{t,j-1}\right)\Big\|_F^2\\
&\quad+4\Big\|\nabla_W F_i\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_i^{t,j-1}\right)-\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)\Big\|_F^2\\
&\quad+4\Big\|\nabla_W F_k\left(\mathbf{W}_k^{t,j-1},\mathbf{H}_k^{t,j-1}\right)-\nabla_W F_k\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_k^{t,j-1}\right)\Big\|_F^2\\
&\quad+4\Big\|\nabla_W F_k\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}_k^{t,j-1}\right)-\nabla_W F\left(\overline{\mathbf{W}}^{t,j-1},\mathbf{H}^{t,j-1}\right)\Big\|_F^2\\
&\overset{(d)}{\le}4\left(L_{W_i}^t\right)^2\Big\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_i^{t,j-1}\Big\|_F^2\\
&\quad+4\left(L_{W_k}^t\right)^2\Big\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_k^{t,j-1}\Big\|_F^2+8\zeta^2 \qquad (73)
\end{aligned}
$$

where (d) follows from Assumption 4. Then, substituting (73) into (72) gives rise to

$$\begin{aligned}
&\sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}(L_{W_i}^t)^2\mathbb{E}\left[\left\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\right\|_F^2\right]\\
&\le \sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}(L_{W_i}^t)^2\left(\frac{(r-Q_1-1)}{K(\eta^t)^2}\sum_{j=Q_1}^{r-2}\sum_{i=1}^{N}\left(4(L_{W_i}^t)^2\left\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_i^{t,j-1}\right\|_F^2\right.\right.\\
&\quad\left.\left.+8\zeta^2+4(L_{W_k}^t)^2\left\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_k^{t,j-1}\right\|_F^2\right)\right)\\
&\overset{(e)}{=}\frac{N}{K}\sum_{r=Q_1+1}^{Q^t}\frac{4(r-Q_1-1)}{\left(\eta^t/L_W^t\right)^2}\sum_{j=Q_1}^{r-2}\sum_{i=1}^{N}(L_{W_i}^t)^2\left\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_i^{t,j-1}\right\|_F^2\\
&\quad+\frac{N}{K}\sum_{r=Q_1+1}^{Q^t}\frac{8(r-Q_1-1)^2}{\left(\eta^t/L_W^t\right)^2}\zeta^2\\
&\quad+\frac{N}{K}\sum_{r=Q_1+1}^{Q^t}\frac{4(r-Q_1-1)}{\left(\eta^t/L_W^t\right)^2}\sum_{j=Q_1}^{r-2}\sum_{i=1}^{N}(L_{W_i}^t)^2\\
&\quad\cdot\left(\frac{L_{W_i}^t}{L_W^t}\right)^2\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_i^{t,j-1}\|_F^2\\
&\overset{(f)}{\le}\frac{2NQ_2^t(Q_2^t-1)}{K\alpha_2^2}\left(1+\frac{\overline{L}_W^2}{\underline{L}_W^2}\right)\sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}(L_{W_i}^t)^2\left\|\overline{\mathbf{W}}^{t,j-1}-\mathbf{W}_i^{t,j-1}\right\|_F^2\\
&\quad+\frac{4NQ_2^t(Q_2^t-1)(2Q_2^t-1)\zeta^2}{3K\alpha_2^2}
\end{aligned}\tag{74}$$

where (e) follows since $(L_W^t)^2=(1/N)\sum_{i=1}^{N}(L_{W_i}^t)^2$; (f) follows due to $[(L_{W_i}^t)^2/(L_W^t)^2]\le[(\overline{L}_W^2)/(\underline{L}_W^2)]$ and $\eta^t=\alpha_2 L_W^t$, and:

$$\begin{aligned}
&\sum_{r=Q_1+1}^{Q^t}(r-1-Q_1)\sum_{j=Q_1}^{r-2}a_j\\
&\quad\le\sum_{r=Q_1+1}^{Q^t}\frac{Q_2^t(Q_2^t-1)}{2}a_{r-1}\quad\forall a_j>0
\end{aligned}\tag{75}$$

and

$$\sum_{r=Q_1+1}^{Q^t}(r-1-Q_1)^2=\frac{Q_2^t(Q_2^t-1)(2Q_2^t-1)}{6}.\tag{76}$$

Since $\alpha_2\ge Q_2^t\sqrt{3(1+\overline{L}_W^2/\underline{L}_W^2)}$, implies $\alpha_2^2\ge 2Q_2^t(Q_2^t-1)(1+\overline{L}_W^2/\underline{L}_W^2)$. After rearranging (74), we obtain

$$\begin{aligned}
&\sum_{r=Q_1+1}^{Q^t}\sum_{i=1}^{N}(L_{W_i}^t)^2\mathbb{E}\left[\|\overline{\mathbf{W}}^{t,r-1}-\mathbf{W}_i^{t,r-1}\|_F^2\right]\\
&\le\frac{4NQ_2^t(Q_2^t-1)(2Q_2^t-1)\zeta^2}{3K\left(\alpha_2^2-2Q_2^t(Q_2^t-1)\left(1+\overline{L}_W^2/\underline{L}_W^2\right)\right)}\\
&\overset{(g)}{\le}\frac{4NQ_2^t(Q_2^t-1)(2Q_2^t-1)\zeta^2}{K\alpha_2^2}\\
&=\frac{4NC_1^t\zeta^2}{K\alpha_2^2}
\end{aligned}\tag{77}$$

where $C_1^t=Q_2^t(Q_2^t-1)(2Q_2^t-1)$, (g) follows since $\alpha_2^2-2Q_2^t(Q_2^t-1)(1+\overline{L}_W^2/\underline{L}_W^2)>\alpha_2^2/3$. ■

APPENDIX E
PER-ITERATION COMPLEXITY OF ALGORITHM 1

According to (17) and (18), let us revisit $\mathbf{H}_i^{t,r}$ and $\mathbf{W}_i^{t,r}$ as follows:

$$\mathbf{H}_i^{t,r}=\left[\mathbf{H}_i^{t,r-1}-\frac{1}{\gamma_i^t}\nabla_{H_i}F_i\left(\mathbf{W}^{t-1},\mathbf{H}_i^{t,r-1}\right)\right]^+\tag{78}$$

$$\mathbf{W}_i^{t,r}=\mathbf{W}_i^{t,r-1}-\frac{1}{\eta^t}\nabla_W F_i\left(\mathbf{W}_i^{t,r-1},\mathbf{H}_i^{t,Q_i};\mathcal{B}_i^{t,r}\right).\tag{79}$$

For simplicity, we omit the outer iteration number t and inner iteration number r. By the definition of $F_i(\mathbf{W},\mathbf{H}_i)$ in (14), $\nabla_{H_i}F_i(\mathbf{W}_i,\mathbf{H}_i)$ and $\nabla_W F_i(\mathbf{W}_i,\mathbf{H}_i;\mathcal{B}_i)$ can be computed as

$$\begin{aligned}
\nabla_{H_i}F_i(\mathbf{W}_i,\mathbf{H}_i)&=2\mathbf{W}_i^T\mathbf{W}_i\mathbf{H}_i-2\mathbf{W}_i^T\mathbf{X}_i\\
&\quad+2\rho\mathbf{1}\mathbf{1}^T\mathbf{H}_i+(\mu_h-\rho)\mathbf{H}_i
\end{aligned}\tag{80}$$

$$\nabla_W F_i(\mathbf{W}_i,\mathbf{H}_i;\mathcal{B}_i)=2\mathbf{W}_i\mathbf{H}_i\mathbf{H}_i^T-2\mathbf{X}_i\mathbf{H}_i^T+\mu_w\mathbf{W}_i.\tag{81}$$

Thus, the complexity order of computing $\mathbf{H}_i^{t,r}$ [mainly due to (80)] at each client i can be estimated as

$$\begin{aligned}
&\left(mk^2+k^2n_i+kn_i\right)+(mkn_i+n_ik)+n_ik^2+2n_ik\\
&\quad\Longrightarrow\ \mathcal{O}\left((m+n_i)k^2+mn_ik\right)
\end{aligned}\tag{82}$$

and that of computing $\mathbf{W}_i^{t,r}$ [mainly due to (81)] as

$$(mkb+mkb+mk)+(mkb+mk)+mk\ \Longrightarrow\ \mathcal{O}(mkb).\tag{83}$$

Because the complexity of $\mathbf{W}_i^{t,r}$ is much smaller than that of $\mathbf{H}_i^{t,r}$ (due to $b=|\mathcal{B}_i|<n_i\ll n=\sum_i^N n_i$), the total complexity order of updating $\mathbf{H}_i^{t,r}$ and $\mathbf{W}_i^{t,r}$ can be approximated by that of updating $\mathbf{H}_i^{t,r}$. As a result, provided that all the N clients (the worst case) join the learning process, one can obtain the total complexity order as $\mathcal{O}((mN+n)k^2+mnk)$ at the client side. Moreover, the complexity order at the PS side is simply $\mathcal{O}(mkN)$.

REFERENCES

[1] Z. Wang, X. Wang, R. Sun, and T.-H. Chang, "Federated semi-supervised learning with class distribution mismatch," 2021, *arXiv:2111.00010*.

[2] H. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," 2016, *arXiv:1602.05629*.

[3] Y. Li, S. Wang, T.-H. Chang, and C.-Y. Chi, "Federated stochastic primal–dual learning with differential privacy," 2022, *arXiv:2204.12284*.

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Areas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1–10.

[5] S. Li, S. Hou, B. Buyukates, and S. Avestimehr, "Secure federated clustering," 2022, *arXiv:2205.15564*.

[6] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5693–5700.

[7] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 2938–2948.

[8] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 937–947.

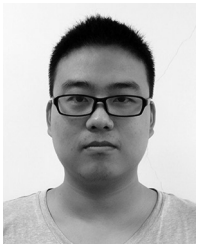
[9] W. Wei and L. Liu, "Gradient leakage attack resilient deep learning," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 303–316, 2021.

[10] Y. Li, T.-H. Chang, and C.-Y. Chi, "Secure federated averaging algorithm with differential privacy," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2020, pp. 1–6.

[11] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[12] S. Wang, T.-H. Chang, Y. Cui, and J.-S. Pang, "Clustering by orthogonal NMF model and non-convex penalty optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 5273–5288, 2021.
[13] M. Stallmann and A. Wilbik, "Towards federated clustering: A federated fuzzy *c*-means algorithm (FFCM)," 2022, *arXiv:2201.07316*.
[14] A. Kolluri, T. Baluta, and P. Saxena, "Private hierarchical clustering in federated networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2021, pp. 2342–2360.
[15] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," 2019, *arXiv:1906.06629*.
[16] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3710–3722, Aug. 2021.
[17] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi, "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 3407–3416.
[18] D. K. Dennis, T. Li, and V. Smith, "Heterogeneity for the win: One-shot federated clustering," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 2611–2620.
[19] J. Ma, G. Long, T. Zhou, J. Jiang, and C. Zhang, "On the convergence of clustered federated learning," 2022, *arXiv:2202.06187*.
[20] H. Ding, Y. Liu, L. Huang, and J. Li, "K-means clustering with distributed dimensions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1339–1348.
[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.*, 2006, pp. 486–503.
[22] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. 13th Annu. ACM/SIAM Symp. Discr. Algorithms*, 2019, pp. 2468–2479.
[23] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 6277–6287.
[24] X. Shen, Y. Liu, and Z. Zhang, "Performance-enhanced federated learning with differential privacy for Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 24079–24094, Dec. 2022.
[25] Y. Li, S. Wang, C.-Y. Chi, and T. Q. Quek, "Differentially private federated learning in edge networks: The perspective of noise reduction," *IEEE Netw.*, vol. 36, no. 5, pp. 167–172, Sep./Oct. 2022.
[26] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Rep. (ICLR)*, 2020, pp. 1–26.
[27] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable *k*-means++," in *Proc. VLDB Endow.*, 2012, pp. 622–633.
[28] T. Kucukyilmaz, "Parallel *k*-means algorithm for shared memory multiprocessors," *J. Comput. Commun.*, vol. 2, no. 11, pp. 15–23, 2014.
[29] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Knowl. Disc. Data Min. (KDD)*, 1996, pp. 226–231.
[30] M.-F. F. Balcan, S. Ehrlich, and Y. Liang, "Distributed *k*-means and *k*-median clustering on general topologies," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 1995–2003.
[31] W. Pedrycz, "Federated FCM: Clustering under privacy requirements," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 8, pp. 3384–3388, Aug. 2022.
[32] E. Hernández-Pereira, O. Fontenla-Romero, B. Guijarro-Berdiñas, and B. Pérez-Sánchez, "Federated learning approach for spectral clustering," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2021, pp. 423–428.
[33] C. Li, G. Li, and P. K. Varshney, "Federated learning with soft clustering," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7773–7782, May 2022.
[34] C. Xu, Y. Qu, Y. Xiang, and L. Gao, "Asynchronous federated learning on heterogeneous devices: A survey," 2021, *arXiv:2109.04269*.
[35] S. Wang and T.-H. Chang, "Federated matrix factorization: Algorithm design and application to data clustering," *IEEE Trans. Signal Process.*, vol. 70, pp. 1625–1640, 2022.
[36] J. Chung, K. Lee, and K. Ramchandran, "Federated unsupervised clustering with generative models," in *Proc. AAAI Int. Workshop Trustable Verifiable Auditable Feder. Learn.*, 2022, pp. 1–9.
[37] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 1175–1191.
[38] R. Mclendon et al., "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.
[39] Y. LeCun, C. Cortes, and C. Burges. "The MNIST database." Online]. Available: http://yann.lecun.com/exdb/mnist
[40] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.
[41] B. Yang, X. Fu, and N. D. Sidiropoulos, "Learning from hidden traits: Joint factor analysis and latent clustering," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 256–269, Jan. 2017.
[42] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. Symp. Discr. Algorithms (SODA)*, 2007, pp. 1027–1035.
[43] T. Li, A. K. Sahu, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2020, pp. 1–12.
[44] S. Wang, T.-H. Chang, Y. Cui, and J.-S. Pang, "Clustering by orthogonal non-negative matrix factorization: A sequential non-convex penalty approach," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 5576–5580.
[45] H. Kim and H. Park, "Sparse non-negative matrix factorization via alternating non-negative-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.
[46] W. E. Zhang, M. Tan, Q. Z. Sheng, L. Yao, and Q. Shi, "Efficient orthogonal non-negative matrix factorization over Stiefel manifold," in *Proc. ACM Int. Conf. Inf. Knowl. Manag. (ICKM)*, 2016, pp. 1743–1752.
[47] K. Yu, S. Yu, and V. Tresp, "Soft clustering on graphs," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 1–8.
[48] J. Kŏnecný, H. B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," in *Proc. NeuIPS Optim. Mach. Learn. Workshop*, 2015, pp. 1–5.
[49] J. Kŏnecný, H. B. McMahan, D. Ramage, and P. Richtarik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.
[50] D. Chai, L. Wang, K. Chen, and Q. Yang, "Secure federated matrix factorization," *IEEE Intell. Syst.*, vol. 36, no. 5, pp. 11–20, Sep./Oct. 2020.
[51] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, Jun. 2001.
[52] 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.
[53] C.-Y. Chi, W.-C. Li, and C.-H. Lin, *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*. Boca Raton, FL, USA: CRC Press, Feb. 2017.
[54] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, pp. 459–494, Aug. 2014.

Yiwei Li (Member, IEEE) is currently pursuing the Ph.D. degree with the National Tsing Hua University, Hsinchu, Taiwan.

He was a visiting Ph.D. student with The Chinese University of Hong Kong, Shenzhen, China, from September 2019 to February 2022, and he served as a Research Assistant with Quanzhou Institute of Equipment Manufacturing, Chinese Academy of Sciences, Quanzhou, Fujian, China, from April 2015 to August 2017. His current primary research interests include security and privacy protection in federated learning, distributed optimization, with applications spanning the Internet of Things, wireless communication system, and intelligent mobile-edge networks.



Shuai Wang received the B.Sc. degree in computer science and technology from Southwest University, Chongqing, China, in 2012, and the Ph.D. degree from the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China, in 2021.

He is currently a Postdoctoral Research Fellow with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore. Before that, he was a Research Assistant with Hong Kong University of Science and Technology, Hong Kong, from August 2014 to August 2015, and with The Hong Kong Polytechnic University, Hong Kong, from October 2015 to July 2016. His current primary research interests include optimization algorithms for signal processing, machine learning and communication systems, distributed optimization and federated learning (FL), such as communication-efficient FL algorithms, data security, and privacy protection in FL.

Chong-Yung Chi (Life Fellow, IEEE) received the B.S. degree in electrical engineering from Tatung Institute of Technology, Taipei, Taiwan, in 1975, the master's degree in electrical engineering from National Taiwan University, Taipei, in 1977, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1983.

He is currently a Professor with National Tsing Hua University, Hsinchu, Taiwan. He has published more than 240 technical papers, including more than 90 journal papers (mostly in IEEE TRANSACTIONS ON SIGNAL PROCESSING), more than 140 peer-reviewed conference papers, and two books, including a textbook, *Convex Optimization for Signal Processing and Communications From Fundamentals to Applications* (CRC Press, 2017) (which has been popularly used in a series of invited intensive short courses at ten top-ranking universities in Mainland China since 2010 before its publication). His current research interests include signal processing for wireless communications and networking, convex analysis and optimization for blind source separation, biomedical and hyperspectral image analysis, graph-based learning and signal processing, and data security and privacy protection in machine learning.

Prof. Chi received the 2018 IEEE Signal Processing Society Best Paper Award, entitled *Outage Constrained Robust Transmit Optimization for Multiuser MISO Downlinks: Tractable Approximations by Conic Optimization*, IEEE TRANSACTIONS ON SIGNAL PROCESSING, vol. 62, no. 21, November 2014. He has been a Technical Program Committee Member for many IEEE-sponsored and co-sponsored workshops, symposiums, and conferences on signal processing and wireless communications, including the Co-Organizer and the General Co-Chairman of 2001 IEEE Workshop on Signal Processing Advances in Wireless Communications. He was an Associate Editor for four IEEE Journals, including the IEEE TRANSACTIONS ON SIGNAL PROCESSING for nine years, from May 2001 to April 2006 and January 2012 to December 2015, and a member of the Signal Processing Theory and Methods Technical Committee from 2005 to 2010, Signal Processing for Communications and Networking Technical Committee from 2011 to 2016, Sensor Array and Multichannel Technical Committee from 2013 to 2018, and IEEE Signal Processing Society. He is an AAIA Fellow.

Tony Q. S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008.

He is currently the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD), Singapore, and an ST Engineering Distinguished Professor, where he also serves as the Director for the Future Communications Research and Development Programme, the Head of the Information Systems Technology and Design Pillar, and the Deputy Director for the SUTD-ZJU IDEA. His current research topics include wireless communications and networking, network intelligence, non-terrestrial networks, open radio access network, and 6G.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards–Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, and the 2022 IEEE Signal Processing Society Best Paper Award. He has been actively involved in organizing and chairing sessions, and has served as a member for the Technical Program Committee as well as the symposium chair for a number of international conferences. He is currently serving as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He is a Fellow of the Academy of Engineering Singapore.