Secure Aggregation for Clustered Federated Learning with Passive Adversaries

Hasin Us Sami Başak Güler

***Abstract*—Clustered federated learning is a popular paradigm to tackle data heterogeneity in federated learning, by training personalized models for groups of users with similar data distributions. A critical challenge is to protect the privacy of individual user updates, as the latter can reveal extensive information about sensitive local datasets. To do so, a recent promising approach is information-theoretic secure aggregation, where parties learn the aggregate (sum) of user updates, but no further information is revealed about the individual updates. In this work, we present the first single-server secure aggregation framework in the context of clustered federated learning, to learn the aggregate of user updates for any clustering of users, but without learning any information about the local updates or cluster identities. Our framework can achieve linear communication complexity under formal information-theoretic privacy guarantees, while providing key trade-offs between communication and computation complexity, adversary tolerance, and resilience to user dropouts.**

***Index Terms*—Clustered federated learning, secure aggregation, distributed learning, coded computing.**

I. Introduction

Federated learning (FL) is a distributed learning framework to train machine learning models over the data stored and processed locally across a large number of wireless devices (users) [1]. Unlike traditional centralized training architectures, where all data is collected by a central party who performs training, FL keeps the data on device. Instead, each user updates the trained model locally on their local data, and then the local updates (e.g., gradients) are aggregated (often by a central server) to form a global model. In doing so, users always keep the data on device, and send only the intermediate computations (e.g., local gradients).

Due to this *on-device* learning architecture (data never leaves the device, but only the local updates are communicated), FL has been highly popular in privacy-sensitive applications, such as healthcare. On the other hand, recent *gradient inversion attacks* have shown that the local updates sent by the users (such as gradients) can still reveal extensive information about the local datasets [2–4]. Secure aggregation (SA) protocols have been introduced to address this challenge, by revealing only the *sum of the local updates* to the server during training, while hiding the contents of individual updates sent from each user using information-theoretic or cryptographic tools [5–12]. In doing so, SA ensures that no further information is revealed beyond the sum of the local updates, preventing the server from associating the aggregated updates with any particular user. SA can further be combined with complementary privacy-preserving mechanisms such as differential privacy [13], [14] and can even benefit the latter [15], [16].

A major challenge of FL is the severe data heterogeneity across the users, which slows down training, and degrades model accuracy [17]. More importantly, training a single model (across the entire network) may disproportionately penalize the performance of underrepresented users [18]. Clustered FL is a recent approach to tackle this challenge by training multiple models, each adapted to a *group of users with similar data distributions* [19–25]. The training process alternates between clustering the users with respect to their data distributions, and training a distinct model within each cluster. For the latter, the server collects and aggregates the local updates (gradients) from the users assigned to each cluster, to update the model designated for that cluster. Several complementary approaches also explore addressing data heterogeneity by designing a personalized model for *each user* through fine-tuning or meta-learning [26–29]. In contrast, clustered FL targets *group-level personalization*, where the server maintains personalized models to serve *groups of users* with similar characteristics, while avoiding excessive memory and storage costs to handle a large number of models.

In this work, our goal is to develop an SA framework for clustered FL. A naive approach is to leverage conventional SA protocols to aggregate the local gradients of the users assigned to each cluster (independently from other clusters). On the other hand, doing so requires the server to learn the cluster identity of each user, which itself is highly sensitive information, revealing which users have similar data distributions [23]. An adversarial user can further infer sensitive information about the characteristics of honest users assigned to the same cluster, simply by leveraging the similarity between the distributions. Importantly, underrepresented users are the most vulnerable to these types of attacks, due to the lack of a large number of honest users with similar data distributions, i.e., same cluster identity. Moreover, clusters may vary throughout the training, using which one may reveal the local gradients by comparing the aggregated updates received at different training rounds [30]. As such, here we ask the following question:

- *How can we enable SA for clustered FL, for the server to learn the aggregate of local gradients for each cluster, but without learning any information about the local gradients or cluster identities of individual users?*

Hasin Us Sami and Başak Güler are with the Department of Electrical and Computer Engineering, University of California, Riverside, CA, 92521 USA (e-mail: hsami003@ucr.edu, bguler@ece.ucr.edu).

An earlier version of this work has been presented at the 2023 IEEE International Symposium on Information Theory (ISIT).

This research was sponsored in part by the OUSD (R&E)/RT&L under Cooperative Agreement Number W911NF-20-2-0267, NSF CAREER Award CCF-2144927, and the UCR OASIS Funding Award.

To address this challenge, in this work we propose the first single-server SA frameworks for clustered FL. In all proposed frameworks, the server can perfectly recover the aggregate of local gradients for each cluster, but without learning any further information about the cluster identities or local gradients of the users. All proposed frameworks ensure strong information-theoretic privacy guarantees, while providing a trade-off between the communication and computation overhead, round complexity, and resilience to user dropouts (e.g., due to poor channel conditions). Our contributions can be summarized as follows:

- We propose SA in the context of clustered FL, where the server aggregates the local gradients from multiple clusters of users simultaneously, without learning any information about the cluster identities or local gradients.
- We propose the first single-server SA framework for clustered FL. By introducing an offline-online trade-off, our framework can achieve a linear online communication complexity, while offloading the communication-intensive operations to a data-agnostic offline phase.
- For all proposed frameworks, we demonstrate the formal information-theoretic privacy guarantees and identify the key performance trade-offs between the communication/computation overhead, privacy against adversaries, round complexity, and resilience to user dropouts.

II. Related Works

For *group-level personalization*, a hierarchical clustering approach is proposed in [19] by partitioning the users into clusters according to the similarity of their local model updates, and a distinct model is trained for each cluster. Reference [20] determines the clusters according to the cosine similarity between different local updates. References [21–23] propose an alternating optimization approach, which alternates between clustering users based on the similarity of their local data distributions, and training a distinct model for each cluster. Reference [24] considers the setting where local datasets are from a mixture of distributions, whereas [25] considers fairness across the clusters. In contrast, [26–29, 31] adopt a *user-level personalization* approach. To this end, [32] adds a proximal term to the local objectives to improve performance in the presence of data heterogeneity. A multi-task learning approach is proposed in [26] to simultaneously tackle data and system heterogeneity. A meta-learning approach is proposed in [27, 28, 31] to provide user-specific models based on local dataset distributions.

Secure aggregation (SA) was introduced in [5, 6], where the local gradients are obfuscated by pairwise additive random masks. The masks cancel out upon aggregation, allowing the server to recover the aggregate of the true gradients. While these works focus on cryptographic security (against adversaries with bounded computational capability), more recent works consider information-theoretic SA, where adversaries have unbounded computational power [7–12]. To this end, [7] proposes a circular aggregation strategy, whereas [8] introduces a one-shot aggregation technique. Reference [9] considers efficient randomness generation with low storage cost, whereas [10] provides a trade-off between communication load and active communication links, and [11] introduces resource-aware SA with quantization.

These works are agnostic to the data heterogeneity across users, and focus on training a single model. Concurrent work [33] considers a two-server secure multi-party computing protocol to aggregate the local updates from different clusters of users. However, unlike SA (which is based on a single-server architecture), this work requires two honest (non-colluding) servers who interact with the users and each other to carry out a secure two-party protocol, but do not share any sensitive information with each other in an attempt to breach user privacy. In contrast, our goal is to develop a single-server secure aggregation framework, to facilitate privacy-preserving training architectures for clustered FL. Compared to the two-server setting, the single server setting carry the additional challenge where the aggregation is handled by a single server, while still being able to keep the individual models and the cluster identities of the users private.

Organization. The remainder of the paper is organized as follows. Section III introduces the system model, Section IV presents our frameworks. Sections V and VI provide the theoretical analysis and experiments, respectively. Section VII discusses extensions to different adversary models, and Section VIII concludes the paper. Throughout the paper, x denotes a scalar, $\mathbf{x}$ is a vector, $\mathbf{X}$ represents a matrix, and $\mathcal{X}$ denotes a set, where $[N]$ is the set $\{1, \ldots, N\}$.

III. Problem Formulation

Clustered FL. We consider a distributed network of N users and a server. The local dataset $\mathcal{D}_i$ of user $i \in [N]$ is realized from one of K distributions denoted by $\mathcal{P}_1, \ldots, \mathcal{P}_K$. The goal is to train K models $\mathbf{w}_1, \ldots, \mathbf{w}_K$, where model $\mathbf{w}_k \in \mathbb{R}^d$ is trained to minimize the loss function,

$$F_k(\mathbf{w}_k) \triangleq \mathbb{E}_{\xi \sim \mathcal{P}_k}[f(\mathbf{w}_k, \xi)] \quad \forall k \in [K], \tag{1}$$

where ξ is a data sample realized from distribution $\mathcal{P}_k$ and $f(\mathbf{w}_k, \xi)$ denotes the stochastic loss function computed on the data sample ξ and model $\mathbf{w}_k$. Then, the optimal model is given by,

$$\mathbf{w}_k^* = \arg\min_{\mathbf{w}_k} F_k(\mathbf{w}_k) \quad \forall k \in [K]. \tag{2}$$

To solve (1), clustered FL [21–23] takes an iterative approach, that alternates between partitioning the users into K clusters with respect to the similarity of the local datasets, and training K global models (one for each cluster). At each iteration t, the server broadcasts the current state of the K global models $\{\mathbf{w}_k(t)\}_{k\in[K]}$ to all users. Then, user $i \in [N]$ computes a local empirical loss,

$$f_i(\mathbf{w}_k(t)) \triangleq \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} f(\mathbf{w}_k(t), \xi) \tag{3}$$

for each model $\{\mathbf{w}_k(t)\}_{k\in[K]}$, and selects the cluster with the minimum loss,

$$c_i^{(t)} \triangleq \arg\min_{k\in[K]} f_i(\mathbf{w}_k(t)). \tag{4}$$

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3366394

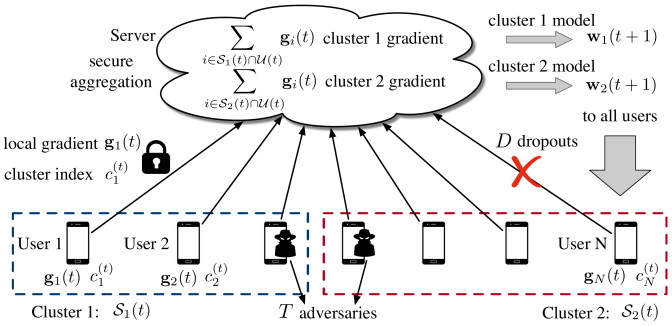


Fig. 1. **Secure aggregation for clustered FL.** The server learns the aggregate of the local gradients $\sum_{i\in\mathcal{S}_k\cap\mathcal{U}} \mathbf{g}_i(t)$ for each cluster $k \in [K]$, *without learning which users belong to which cluster*, or the local gradients $\mathbf{g}_i(t)$ of the individual users.

Next, user $i \in [N]$ computes a local gradient for the model of the selected cluster,

$$\mathbf{g}_i(t) \triangleq \nabla f_i(\mathbf{w}_{c_i^{(t)}}(t)) \tag{5}$$

and sends the local gradient from (5), along with the cluster index from (4), to the server. Then, the server updates the global model for each cluster, by aggregating the local gradients received from users assigned to that cluster,

$$\mathbf{w}_k(t+1) = \mathbf{w}_k(t) - \frac{\eta}{N} \sum_{i\in\mathcal{S}_k(t)\cap\mathcal{U}(t)} \mathbf{g}_i(t) \quad \forall k \in [K], \tag{6}$$

where η is the learning rate, and $\mathcal{S}_k(t) \triangleq \{i : c_i^{(t)} = k, i \in [N]\}$ denotes the set of users assigned to cluster k at iteration t. At each training iteration, up to D out of N users may drop out from the protocol due to various reasons, such as poor channel conditions or low battery. Accordingly, $\mathcal{U}(t) \subseteq [N]$ denotes the set of surviving users at iteration t, who successfully send their local gradient $\mathbf{g}_i(t)$ to the server, where $|\mathcal{U}(t)| \geq N - D$.

Remark 1. *The key intuition behind the clustered learning mechanism is that when user datasets are sampled from K different distributions, the optimal model for each distribution should minimize the local loss for the corresponding users [23]. Accordingly, at each training round, the clustering mechanism identifies the group of users for which a given (global) model performs the best, and then further updates the model using the local datasets of the corresponding users.*

Threat model. We consider an honest-but-curious (passive) adversary model (as is the most common threat model in SA), where adversaries follow the protocol, but try to reveal additional information about the local datasets of honest users from the messages exchanged during training [7], [9]. Out of N users, any set of up to T users can be adversarial. Adversarial users may collude with each other, and the adversaries from one cluster may collude with the adversaries from different clusters. The server is also honest-but-curious and may collude with the adversarial users.

Information-theoretic secure aggregation. Our goal is to enable the server to compute the sum of the local gradients $\sum_{i\in\mathcal{S}_k(t)\cap\mathcal{U}(t)} \mathbf{g}_i(t)$ for each cluster $k \in [K]$, in order to update the model from (6) correctly, but without learning any further information about the local gradients or the cluster identities of the users. Formally, this condition can be stated as follows:

$$I\Big(\{\mathbf{g}_i(t), c_i^{(t)}\}_{[N]\setminus\mathcal{T}}; \mathcal{M}_\mathcal{T} \Big| \Big\{\sum_{i\in\mathcal{S}_k(t)\cap\mathcal{U}(t)} \mathbf{g}_i(t)\Big\}_{k\in[K]}, \{\mathbf{g}_i(t), c_i^{(t)}\}_{i\in\mathcal{T}}, \mathcal{G}_\mathcal{T}\Big) = 0 \tag{7}$$

for any set of adversarial users $\mathcal{T}$ such that $|\mathcal{T}| \leq T$, where $\mathcal{M}_\mathcal{T}$ denotes the collection of all messages received by the adversaries and the server, and $\mathcal{G}_\mathcal{T}$ is the set of randomness generated by the adversaries during training. We then ask the following question:

- How can the server compute the aggregate of local gradients from (6) for all K clusters, under the information-theoretic privacy guarantees from (7)?

To address this challenge, in this work we propose three SA protocols, with different trade-offs in terms of communication/computation overhead, round complexity, and dropout resilience. Similar to [5, 7, 9], our frameworks are bound to finite field computations, where each user converts their local gradient $\mathbf{g}_i(t) \in \mathbb{R}^d$ from the real domain to a finite field $\mathbb{F}_q$ of integers modulo a large prime q. The details of this conversion is provided in Appendix A. In the sequel, $\overline{\mathbf{g}}_i(t) \in \mathbb{F}_q^d$ denotes the finite field representation of $\mathbf{g}_i(t)$. All computations are then performed in $\mathbb{F}_q$. Our system model is illustrated in Fig. 1. Similar to [7, 10, 34], we assume that there exists direct (peer-to-peer) communication links between the users, in addition to the user-to-server links. In scenarios where peer-to-peer links are not available, one can utilize cryptographic encryption mechanisms to forward all messages through the server [5, 6].

We next present the details of our frameworks.

IV. Clustered Secure Aggregation

We next present three approaches to SA for clustered FL. For notational clarity, we omit the iteration index t in our exposition. In all frameworks, a new set of randomness is generated at each training round. The randomness generation in the offline phases can be carried out when the network load is low, or can be overlapped with other components of training.

A. Clustered Secret Gradient Sharing (CSGS)

In our first framework, users encode their local gradients by partitioning them into multiple shards, and combining them with T random masks. Then, each user sends an encoded gradient to every other user. The random masks hide the true gradient and cluster identity against up to T adversaries, while the encoding mechanism provides a trade-off between communication complexity and resilience to user dropouts. We next describe the details of this procedure.

Initially, the server generates N distinct public parameters $\alpha_1, \ldots, \alpha_N$ independently and uniformly at random (without replacement) from $\mathbb{F}_q$, and sends them to the users prior to training. Then, each user $i \in [N]$ partitions its local gradient $\overline{\mathbf{g}}_i$ into L equal-sized shards,

$$\overline{\mathbf{g}}_i = \begin{bmatrix}\overline{\mathbf{g}}_{i1}^\mathrm{T} & \cdots & \overline{\mathbf{g}}_{iL}^\mathrm{T}\end{bmatrix}^\mathrm{T}, \tag{8}$$

and generates T independent (uniformly) random vectors $\mathbf{v}_{i1},\ldots,\mathbf{v}_{iT}\in\mathbb{F}_q^{\frac{d}{L}}$. Then, user i forms a degree $KL+T-1$ polynomial,

$$f_i(\alpha)\triangleq\sum_{l=1}^{L}\alpha^{(c_i-1)L+l-1}\overline{\mathbf{g}}_{il}+\sum_{l=1}^{T}\alpha^{KL+l-1}\mathbf{v}_{il},\quad(9)$$

and sends to each user $j\in[N]$ a *coded gradient*,

$$\widetilde{\mathbf{g}}_{ij}\triangleq f_i(\alpha_j).\quad(10)$$

In doing so, some users may drop out from the protocol and fail to send their coded gradients. We denote the set of surviving users at the end of this stage (i.e., users who successfully send their coded gradients from (10)) by $\mathcal{U}_1\subseteq[N]$. To recover the *aggregate* of the local gradients of these *surviving* users, the server then requests the aggregate of the coded gradients,

$$\widetilde{\mathbf{g}}_i\triangleq\sum_{j\in\mathcal{U}_1}\widetilde{\mathbf{g}}_{ji}\quad(11)$$

from each user $i\in[N]$. Note that the computations from (11) can be viewed as evaluations of a degree $KL+T-1$ polynomial,

$$f(\alpha)\triangleq\sum_{j\in\mathcal{U}_1}f_j(\alpha)=\sum_{k\in[K]}\sum_{l\in[L]}\alpha^{(k-1)L+l-1}\Big(\sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{jl}\Big)+\sum_{l=1}^{T}\alpha^{KL+l-1}\Big(\sum_{j\in\mathcal{U}_1}\mathbf{v}_{jl}\Big)\quad(12)$$

at an interpolation point $\alpha=\alpha_i$, where $\widetilde{\mathbf{g}}_i=f(\alpha_i)$. The set of surviving users at the end of this stage (i.e., users who successfully send the sum of the coded gradients in (11)) is defined as $\mathcal{U}_2$, where $\mathcal{U}_2\subseteq\mathcal{U}_1\subseteq[N]$. Since any polynomial f of degree $\deg f$ can be uniquely reconstructed from at least $\deg f+1$ evaluation points, upon receiving the evaluations (11) from the users in $\mathcal{U}_2$, where $|\mathcal{U}_2|\geq KL+T$, the server can reconstruct the aggregate of the local gradients,

$$\sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_j=\begin{bmatrix}\sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{j1}^{\mathrm{T}} & \cdots & \sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{jL}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}},\quad(13)$$

for each cluster $k\in[K]$, using polynomial interpolation. Parameter L controls a trade-off between communication complexity and resilience to user dropouts. Specially, as will be detailed in Section V, the communication overhead is $O(\frac{dN}{L})$ per user, which is inversely proportional to L, whereas the maximum number of user dropouts that can be tolerated is given as $D\leq N-(KL+T)$, which increases by selecting a smaller L.

B. Clustered Masked Gradient Aggregation (CMGA)

Our second framework builds on an online-offline trade-off, by dividing the communication into online (data-dependent) and offline (data-agnostic) phases. The former depends on the datasets, hence can only be carried out after training starts. The latter is independent from data (such as randomness generation), and can be carried out flexibly in advance when the network load is low (accordingly, we assume that the user dropouts occur in the online phase.). The key intuition is then to transfer the intensive communication overhead incurred by large N to the offline phase, by increasing the number of communication rounds. As demonstrated next, one can achieve an online communication overhead of $O(dK)$ (independent from the number of users) while keeping the offline overhead as $O(\frac{dN}{L})$. We next describe the details of the offline and online phases, respectively.

Offline. In the offline phase, the server generates N distinct public parameters $\alpha_1,\ldots,\alpha_N$ independently and uniformly at random (without replacement) from $\mathbb{F}_q$, and sends them to the users. User $i\in[N]$ then generates K random masks $\{\mathbf{r}_{ik}\}_{k\in[K]}$ of size d uniformly at random from $\mathbb{F}_q$, and partitions each mask into L equal-sized shards,

$$\mathbf{r}_{ik}=\begin{bmatrix}\mathbf{r}_{ik1}^{\mathrm{T}} & \cdots & \mathbf{r}_{ikL}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}.\quad(14)$$

Using the random masks generated, user i constructs a polynomial of degree $KL+T-1$,

$$f_i(\alpha)\triangleq\sum_{k=1}^{K}\sum_{l=1}^{L}\alpha^{(k-1)L+l-1}\mathbf{r}_{ikl}+\sum_{l=1}^{T}\alpha^{KL+l-1}\mathbf{v}_{il},\quad(15)$$

where $\mathbf{v}_{il}\in\mathbb{F}_q^{\frac{d}{L}}$ are generated uniformly at random for all $l\in[T]$, and sends an *encoded mask*,

$$\widetilde{\mathbf{r}}_{ij}\triangleq f_i(\alpha_j)\quad(16)$$

to each user $j\in[N]$. The random masks $\{\mathbf{r}_{ik}\}_{k\in[K]}$ will be utilized to hide the true content of the local gradients in the online phase, whereas the random vectors $\{\mathbf{v}_{il}\}_{l\in[T]}$ will hide the *true value of the masks* against up to T adversaries.

Online. In the online phase, each user $i\in[N]$ sends to the server a *masked gradient*,

$$\mathbf{x}_{ik}\triangleq\begin{cases}\overline{\mathbf{g}}_i+\mathbf{r}_{ik} & \text{if } i\in\mathcal{S}_k\\ \mathbf{r}_{ik} & \text{otherwise}\end{cases},\quad(17)$$

for each cluster $k\in[K]$. We define $\mathcal{U}_1$ to represent the set of users who successfully send their masked gradient from (17) to the server. Then, the server aggregates the received masked gradients $\{\mathbf{x}_{ik}\}_{i\in\mathcal{U}_1}$ from the surviving users $\mathcal{U}_1$, by evaluating the sum $\sum_{i\in\mathcal{U}_1}\mathbf{x}_{ik}$ for each cluster $k\in[K]$. On the other hand, to recover the aggregate of the *true gradients* $\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i$ from the masked gradients $\sum_{i\in\mathcal{U}}\mathbf{x}_{ik}$, the server has to remove the aggregate of the *random masks* $\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}$ from the latter. To do so, the server requests the aggregate of the *coded masks*,

$$\widetilde{\mathbf{r}}_i\triangleq\sum_{j\in\mathcal{U}_1}\widetilde{\mathbf{r}}_{ji}\quad(18)$$

from each user $i\in\mathcal{U}_1$. The computations from (18) can be viewed as evaluations of a degree $KL+T-1$ polynomial,

$$f(\alpha)\triangleq\sum_{j\in\mathcal{U}_1}f_j(\alpha)=\sum_{k=1}^{K}\sum_{l=1}^{L}\alpha^{(k-1)L+l-1}\Big(\sum_{j\in\mathcal{U}_1}\mathbf{r}_{jkl}\Big)+\sum_{l=1}^{T}\alpha^{KL+l-1}\Big(\sum_{j\in\mathcal{U}_1}\mathbf{v}_{jl}\Big)\quad(19)$$

at an interpolation point $\alpha=\alpha_i$, where $\widetilde{\mathbf{r}}_i=f(\alpha_i)$. We let $\mathcal{U}_2$ denote the set of users who successfully send the aggregate of the coded masks in (18) to the server, where $\mathcal{U}_2\subseteq\mathcal{U}_1\subseteq[N]$.

Then, upon receiving the evaluations in (18) from any set of at least $KL+T$ users, the server can reconstruct the aggregate of the random masks,

$$\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}=\left[\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik1}^{\mathrm{T}}\cdots\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ikL}^{\mathrm{T}}\right]^{\mathrm{T}}\text{ for }k\in[K]\quad(20)$$

via polynomial interpolation. Then, the server can recover the aggregate of the *true gradients for each cluster*, by removing the random masks in (20) from the masked gradients $\sum_{i\in\mathcal{U}_1}\mathbf{x}_{ik}$ as,

$$\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i=\sum_{i\in\mathcal{U}_1}\mathbf{x}_{ik}-\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\text{ for }k\in[K].\quad(21)$$

CMGA achieves a per-user online communication overhead of $O(dK)$, by offloading the $O(\frac{dN}{L})$ (online) overhead of CSGS to the offline phase, while providing equal resilience against user dropouts $D\leq N-(KL+T)$. On the other hand, when the number of clusters K is large, as is often the case in highly heterogeneous networks, the $O(dK)$ overhead is still significant. Our next framework overcomes this challenge by reducing the online overhead to $O(d+K)$, achieving a linear communication complexity in both the model size d and the number of clusters K, by trading-off the communication overhead with tolerance to user dropouts.

C. Secure Aggregation with Masked Clusters (SAMC)

Our last framework also builds on an online/offline trade-off, where we offload the communication intensive operations to the offline phase. On the other hand, instead of aggregating the masked gradients for each cluster, each user now sends a *one-shot masked gradient* along with a *masked cluster identity*. The two are then combined with encoded random masks generated in the offline phase, in a way that the server can correctly recover the sum of the true gradients for each cluster, without learning any information about their true value. We next describe the details of the offline and online phases.

Offline. In this phase, users generate three Lagrange interpolation polynomials, where the first two will be used to mask the local gradients and cluster identities in the online phase, while the third one will be used to ensure information theoretic privacy during the final reconstruction of the sum of local gradients. Initially, the server generates $2(N+KL+T)-1$ distinct public parameters $\{\alpha_i\}_{i\in[N]}$, $\{\beta_m\}_{m\in[KL+T]}$, $\{\theta_m\}_{m\in\{KL+1,\ldots,2(KL+T-1)+1\}}$, and $\{\lambda_m\}_{m\in[N-T]}$ independently and uniformly at random (without replacement) from $\mathbb{F}_q$, and sends them to the users. Next, each user $i\in[N]$ generates a random mask,

$$\mathbf{r}_i\triangleq\left[\mathbf{r}_{i1}^{\mathrm{T}}\quad\cdots\quad\mathbf{r}_{iL}^{\mathrm{T}}\right]^{\mathrm{T}},\quad(22)$$

where $\mathbf{r}_{il}\in\mathbb{F}_q^{\frac{d}{L}}$ for all $l\in[L]$ are generated uniformly at random (and independently from other elements), and then forms a Lagrange polynomial of degree $KL+T-1$,

$$f_i(\alpha)\triangleq\sum_{l\in[L]}\mathbf{r}_{il}\sum_{k\in[K]}\prod_{\substack{m\in[KL+T]\\\setminus\{(k-1)L+l\}}}\frac{\alpha-\beta_m}{\beta_{(k-1)L+l}-\beta_m}+\sum_{l=KL+1}^{KL+T}\mathbf{v}_{il}\prod_{m\in[KL+T]\setminus\{l\}}\frac{\alpha-\beta_m}{\beta_l-\beta_m},\quad(23)$$

where $\mathbf{v}_{il}\in\mathbb{F}_q^{\frac{d}{L}}$ are uniformly random vectors for all $l\in\{KL+1,\ldots,KL+T\}$, where each element is generated independently from the other elements. Then, user i sends an *encoded mask*,

$$\widetilde{\mathbf{r}}_{ij}\triangleq f_i(\alpha_j)\quad(24)$$

to each user $j\in[N]$. In addition, user i generates K random masks $z_{i1},\ldots,z_{iK}\in\mathbb{F}_q$ (uniformly at random), forms a second Lagrange polynomial of degree $KL+T-1$,

$$h_i(\alpha)\triangleq\sum_{k\in[K]}z_{ik}\sum_{l\in[L]}\prod_{\substack{m\in[KL+T]\\\setminus\{(k-1)L+l\}}}\frac{\alpha-\beta_m}{\beta_{(k-1)L+l}-\beta_m}+\sum_{l=KL+1}^{KL+T}u_{il}\prod_{m\in[KL+T]\setminus\{l\}}\frac{\alpha-\beta_m}{\beta_l-\beta_m},\quad(25)$$

where $u_{il}\in\mathbb{F}_q$ are generated uniformly at random for all $l\in\{KL+1,\ldots,KL+T\}$, and sends an encoded mask,

$$\widetilde{z}_{ij}\triangleq h_i(\alpha_j)\quad(26)$$

to user $j\in[N]$. Finally, user i generates a third Lagrange polynomial of degree $2(KL+T-1)$,

$$v_i(\alpha)\triangleq\sum_{l=KL+1}^{2(KL+T-1)+1}\mathbf{n}_{il}\prod_{m\in[2(KL+T-1)+1]\setminus\{l\}}\frac{\alpha-\theta_m}{\theta_l-\theta_m}\quad(27)$$

where $\theta_l\triangleq\beta_l$ for $l\in[KL]$, and $\mathbf{n}_{il}$ is a random vector of size $\frac{d}{L(N-T)}$ for $l\in\{KL+1,\ldots,2(KL+T-1)+1\}$, where each element is generated independently and uniformly at random from $\mathbb{F}_q$. User i then sends an encoded vector,

$$\widetilde{\mathbf{n}}_{ij}\triangleq v_i(\alpha_j)\quad(28)$$

to user $j\in[N]$. After receiving $\{\widetilde{\mathbf{n}}_{ji}\}_{j\in[N]}$, user i computes,

$$\widetilde{\mathbf{n}}_i\triangleq\left[\sum_{j\in[N]}\lambda_1^{j-1}\widetilde{\mathbf{n}}_{ji}^{\mathrm{T}}\quad\cdots\quad\sum_{j\in[N]}\lambda_{N-T}^{j-1}\widetilde{\mathbf{n}}_{ji}^{\mathrm{T}}\right]^{\mathrm{T}}\quad(29)$$

which can be viewed as evaluations of a Lagrange polynomial $v(\alpha)$ of degree $2(KL+T-1)$,

$$v(\alpha)\triangleq\sum_{l=KL+1}^{2(KL+T-1)+1}\mathbf{n}_l\prod_{m\in[2(KL+T-1)+1]\setminus\{l\}}\frac{\alpha-\theta_m}{\theta_l-\theta_m}\quad(30)$$

such that the computation at user $i\in[N]$ is given by $\widetilde{\mathbf{n}}_i=v(\alpha_i)$, whereas $v(\theta_l)=\mathbf{0}$ for all $l\in[KL]$. Hence, the first KL coefficients are equal to 0, and

$$v(\theta_l)=\mathbf{n}_l=\left[\sum_{j\in[N]}\lambda_1^{j-1}\mathbf{n}_{jl}^{\mathrm{T}}\quad\cdots\quad\sum_{j\in[N]}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}^{\mathrm{T}}\right]^{\mathrm{T}}\quad(31)$$

for all $l\in\{KL+1,\ldots,2(KL+T-1)+1\}$.

Online. In the online phase, each user $i\in[N]$ initially broadcasts a masked local gradient,

$$\mathbf{x}_i\triangleq\overline{\mathbf{g}}_i-\mathbf{r}_i\quad(32)$$

along with a masked cluster index for each cluster $k\in[K]$,

$$y_{ik}\triangleq b_{ik}-z_{ik},\quad(33)$$

where b_{ik} is a binary indicator variable,

$$b_{ik}\triangleq\begin{cases}1 & \text{if } i\in\mathcal{S}_k\\ 0 & \text{otherwise}\end{cases}\quad(34)$$

representing whether user i is assigned to cluster $k \in [K]$. Let $\mathcal{U}_1 \subseteq [N]$ denote the set of surviving users at the end of this stage, i.e., users who successfully send their masked local gradient and cluster index from (32) and (33). To reconstruct the aggregate of local gradients, the server requests from the surviving users $i \in \mathcal{U}_1$,

$$\widetilde{\mathbf{a}}_i \triangleq \sum_{j\in\mathcal{U}_1} \Big(\sum_{k\in[K]} y_{jk} \sum_{l\in[L]} \prod_{\substack{m\in[KL+T]\\ \setminus\{(k-1)L+l\}}} \frac{\alpha_i-\beta_m}{\beta_{(k-1)L+l}-\beta_m} + \widetilde{z}_{ji} \Big)$$
$$\times \Big(\sum_{l\in[L]} \mathbf{x}_{jl} \sum_{k\in[K]} \prod_{\substack{m\in[KL+T]\\ \setminus\{(k-1)L+l\}}} \frac{\alpha_i-\beta_m}{\beta_{(k-1)L+l}-\beta_m} + \widetilde{\mathbf{r}}_{ji} \Big) - \widetilde{\mathbf{n}}_i \quad (35)$$

where the masked gradient $\mathbf{x}_j = \begin{bmatrix} \mathbf{x}_{j1}^{\mathrm{T}} & \cdots & \mathbf{x}_{jL}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ is partitioned into L equal-sized shards. The computations from (35) can be viewed as evaluations of a degree $2(KL+T-1)$ polynomial,

$$f(\alpha) \triangleq \Big(\sum_{j\in\mathcal{U}_1} \phi_j(\alpha)\psi_j(\alpha) \Big) - v(\alpha) \quad (36)$$

such that $\widetilde{\mathbf{a}}_i = f(\alpha_i)$, and

$$\phi_j(\alpha) \triangleq \sum_{k\in[K]} b_{jk} \sum_{l\in[L]} \prod_{\substack{m\in[KL+T]\\ \setminus\{(k-1)L+l\}}} \frac{\alpha-\beta_m}{\beta_{(k-1)L+l}-\beta_m} + \sum_{l=KL+1}^{KL+T} u_{jl} \prod_{m\in[KL+T]\setminus\{l\}} \frac{\alpha-\beta_m}{\beta_l-\beta_m}, \quad (37)$$

$$\psi_j(\alpha) \triangleq \sum_{l\in[L]} \overline{\mathbf{g}}_{jl} \sum_{k\in[K]} \prod_{\substack{m\in[KL+T]\\ \setminus\{(k-1)L+l\}}} \frac{\alpha-\beta_m}{\beta_{(k-1)L+l}-\beta_m} + \sum_{l=KL+1}^{KL+T} \mathbf{v}_{jl} \prod_{m\in[KL+T]\setminus\{l\}} \frac{\alpha-\beta_m}{\beta_l-\beta_m}, \quad (38)$$

where $\overline{\mathbf{g}}_j = \begin{bmatrix} \overline{\mathbf{g}}_{j1}^{\mathrm{T}} & \cdots & \overline{\mathbf{g}}_{jL}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ denotes the local gradient of user j partitioned into L equal-sized shards. We denote the set of surviving users who successfully send their local computation from (35) to the server as $\mathcal{U}_2$, where $\mathcal{U}_2 \subseteq \mathcal{U}_1 \subseteq [N]$. Since $f(\beta_{(k-1)L+l}) = \sum_{j\in\mathcal{U}_1} b_{jk}\overline{\mathbf{g}}_{jl} = \sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1} \overline{\mathbf{g}}_{jl}$ correspond to the *true sum of the local gradients* for each cluster $k \in [K]$ and shard $l \in [L]$, after receiving the local computations (35) from any set of at least $2(KL+T-1)+1$ users, the server can reconstruct $f(\alpha)$ through polynomial interpolation, and recover the sum,

$$\sum_{j\in\mathcal{S}_k\cap\mathcal{U}_1} \overline{\mathbf{g}}_j = \begin{bmatrix} f(\beta_{(k-1)L+1})^{\mathrm{T}} & \cdots & f(\beta_{(k-1)L+L})^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \quad (39)$$

of the local gradients for each cluster $k \in [K]$.

Remark 2. *SAMC reduces the per-user online communication overhead to $O(d+K)$ (down from the $O(Kd)$ overhead of CMGA), while keeping the offline overhead the same. This is achieved by a trade-off between communication overhead and dropout resilience; SAMC slashes the online communication complexity, while requiring a larger number of surviving users for correct recovery of aggregated gradients. A comparison of the communication complexity and dropout resilience of the three frameworks is given in Table I, which will be further detailed in Section V.*

TABLE I
COMPARISON OF COMMUNICATION COMPLEXITY (PER-USER) AND DROPOUT RESILIENCE (MAXIMUM NUMBER OF USER DROPOUTS) FOR THE THREE FRAMEWORKS.

Communication complexity			Dropout resilience
CSGS	online	$O(dN/L)$	$D \le N-(KL+T)$
	offline	–	
CMGA	online	$O(dK)$	$D \le N-(KL+T)$
	offline	$O(dN/L)$	
SAMC	online	$O(d+K)$	$D \le N-2(KL+T)+1$
	offline	$O(dN/L)$	

Remark 3. *The key intuition behind the polynomial $v(\alpha)$ in* (36) *is to ensure privacy during the reconstruction of the final outcomes by the server. Since $v(\beta_{(k-1)L+l}) = 0$ for all $k \in [K], l \in [L]$, in principle, the final outcomes in* (39) *can be recovered by interpolating the polynomial $\sum_{j\in\mathcal{U}_1} \phi_j(\alpha)\psi_j(\alpha)$ directly, by collecting the evaluations $\sum_{j\in\mathcal{U}_1} \phi_j(\alpha_i)\psi_j(\alpha_i)$ from the users, however, additional information may be leaked (beyond the desired outcomes) from the intermediate polynomial coefficients. The masking with $\widetilde{\mathbf{n}}_i = v(\alpha_i)$ prevents such information leakage, as will be demonstrated in Theorem 4.*

V. Theoretical Analysis

We first analyze the per-user communication/computation complexity, privacy against adversaries, and resilience to user dropouts. The dropout resilience of a given framework is quantified by the *recovery threshold*, defined as the minimum number of surviving users required for correct recovery of the aggregate of local gradients.

Theorem 1. *CSGS has a per-user communication complexity $O(\frac{dN}{L})$, per-user computation complexity $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$, and a recovery threshold of $N-D \ge KL+T$.*

Proof. (Communication) The per-user communication overhead consists of: 1) $O(\frac{dN}{L})$ for sending the encoded gradient from (10) to N users, 2) $O(\frac{d}{L})$ for sending (11) to the server.

(Computation) Interpolating a polynomial of degree κ, and evaluating it at κ points has a computational complexity of $\kappa\log^2\kappa\log\log\kappa$ [35]. Then, the per-user computation overhead consists of: 1) $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$ for evaluating the polynomial of degree $KL+T-1$ from (10) at N evaluation points, 2) $O(|\mathcal{U}_1|\frac{d}{L})$ for aggregating the coded vectors received from the surviving users in (11).

(Recovery threshold) To recover the aggregate of the local gradients from (13), the server has to reconstruct the degree $KL+T-1$ polynomial $f(\alpha)$ from (12), which requires the evaluations from any set of at least $KL+T$ surviving users, leading to a recovery threshold $N-D \ge KL+T$. □

Theorem 2. *CMGA has a per-user communication complexity of $O(dK)$ online and $O(\frac{dN}{L})$ offline, per-user computation complexity of $O(\frac{dN}{L})$ online and $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$ offline, and a recovery threshold of $N-D \ge KL+T$.*

Proof. (Communication) The per-user communication overhead consists of the following components. (Online): 1) $O(dK)$ for sending (17) to the server, 2) $O(\frac{d}{L})$ for sending (18) to the server. (Offline): $O(\frac{dN}{L})$ for sending the coded masks from (16) to N users.

(Computation) The per-user computation overhead consists of the following components. (Online): 1) $O(d)$ for computing the masked gradient in (17), 2) $O(|\mathcal{U}_1|\frac{d}{L})$ for aggregating the masks in (18). (Offline): $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$ to evaluate the degree $KL+T-1$ polynomial from (15) at N points.

(Recovery threshold) To recover the aggregated gradients, the server interpolates the degree $KL+T-1$ polynomial $f(\alpha)$ from (19), which requires evaluations from $N-D \geq KL+T$ surviving users. □

Theorem 3. *SAMC has a per-user communication complexity of $O(d+K)$ online and $O(\frac{dN}{L})$ offline, along with a per-user computational complexity $O(N(K+d))$ online and $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$ offline, and a recovery threshold of $N-D \geq 2(KL+T)-1$.*

Proof. (Communication) The per-user communication overhead consists of the following. (Online): 1) $O(d)$ for broadcasting the masked gradient (32), 2) $O(K)$ for broadcasting the masked cluster identity (33), 3) $O(\frac{d}{L})$ for sending (35) to the server. (Offline): 1) $O(\frac{dN}{L})$ for sending (24) to N users, 2) $O(N)$ for sending (26) to N users, 3) $O(\frac{dN}{L(N-T)})$ for sending (28) to N users.

(Computation) The per-user computation overhead consists of the following components. (Online): 1) $O(d)$ for computing the masked local gradient from (32), 2) $O(K)$ for computing the masked cluster assignments in (33), 3) $O(|\mathcal{U}_1|(K+d))$ for computing (35). (Offline): 1) $O(\frac{dN}{L}\log^2(KL+T)\log\log(KL+T))$ for evaluating the polynomial $f_i(\alpha)$ of degree $KL+T-1$ from (23) at N evaluation points, 2) $O(N\log^2(KL+T)\log\log(KL+T))$ for evaluating the polynomial $h_i(\alpha)$ of degree $KL+T-1$ from (25) at N points, 3) $O(\frac{dN}{L(N-T)}\log^2(KL+T)\log\log(KL+T))$ for evaluating the polynomial $v_i(\alpha)$ of degree $2(KL+T-1)$ from (27) at N points, 4) $O(\frac{dN}{L})$ for computing $\widetilde{\mathbf{n}}_i$ from (29).

(Recovery threshold) To aggregate the local gradients, the server interpolates the degree $2(KL+T-1)$ polynomial from (36), using the evaluations (35) of $N-D \geq 2(KL+T-1)+1$ surviving users. □

Remark 4. *The three frameworks provide a trade-off between the online/offline communication complexity, computation cost, and recovery threshold. CMGA reduces the online communication overhead of CSGS by introducing an offline phase. SAMC reduces the online communication by a factor of K compared to CMGA, while increasing the recovery threshold by a constant factor.*

Remark 5. *There is a fundamental trade-off between the privacy against adversaries (T) and dropout resilience (D) in a given network of size N, characterized by the recovery threshold, where one has to decrease T in order to increase D (and vice versa). For both CSGS and CMGA, the maximum number of user dropouts that can be tolerated is given by $D \leq N-KL-T$ from the recovery threshold. As a result, increasing the adversary tolerance T by 1 comes at a cost of reduced dropout resilience D by 1. On the other hand, for SAMC, the maximum number of user dropouts that can be tolerated is $D \leq N+1-2KL-2T$, hence increasing the adversary tolerance T by 1 comes at a cost of reducing the dropout resilience D by 2.*

We next demonstrate the information-theoretic privacy guarantees from (7) for all the three frameworks.

Theorem 4. *(Information-theoretic privacy) All three frameworks CSGS, CMGA, and SAMC provide information-theoretic privacy guarantees from (7) against any set $\mathcal{T}$ of up to $|\mathcal{T}| \leq T$ adversarial users,*

$$I\left(\{\overline{\mathbf{g}}_i, c_i\}_{[N]\setminus\mathcal{T}}; \mathcal{M}_{\mathcal{T}} \middle| \left\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1} \overline{\mathbf{g}}_i\right\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, c_i\}_{i\in\mathcal{T}}, \mathcal{G}_{\mathcal{T}}\right) = 0 \tag{40}$$

where $\mathcal{M}_{\mathcal{T}}$ denotes the collection of all messages received by the server and adversarial users during the protocol, and $\mathcal{G}_{\mathcal{T}}$ is the set of randomness generated by the adversarial users.

Proof. The proof is provided in Appendix B. □

Remark 6. *An upper bound on the mutual information between the local dataset of user $i \in \mathcal{S}_k(t)$ and the gradient aggregate for cluster $k \in [K]$ can be obtained from [36, Theorem 1] as,*

$$I\left(\mathcal{D}_i; \sum_{i\in\mathcal{S}_k(t)\cap\mathcal{U}_1(t)} \overline{\mathbf{g}}_i(t) \middle| \left\{\sum_{i\in\mathcal{S}_k(j)\cap\mathcal{U}_1(j)} \overline{\mathbf{g}}_i(j)\right\}_{j\in[t-1]}\right) \leq O\left(\frac{1}{N_k(t)B}\right)$$

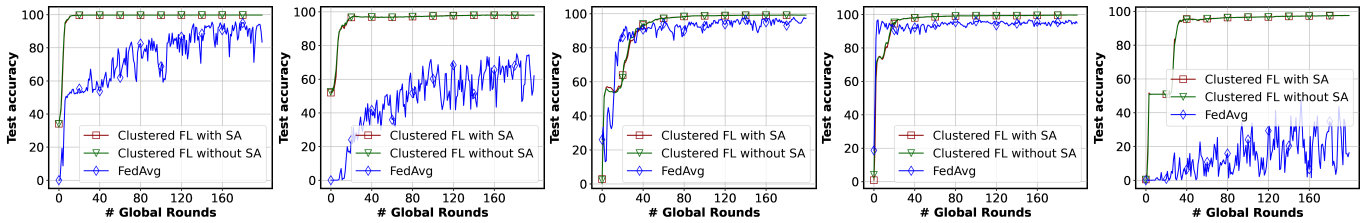
where $N_k(t) \triangleq |\mathcal{S}_k(t)\cap\mathcal{U}_1(t)|$, and B is the batch size for local training at the users. Accordingly, a larger cluster size (i.e., larger number of users within a cluster) reduces the information leakage from the aggregated gradients.

VI. Experiments

In this section, we evaluate the performance of our frameworks with respect to key performance measures; communication overhead, dropout tolerance, and model accuracy.

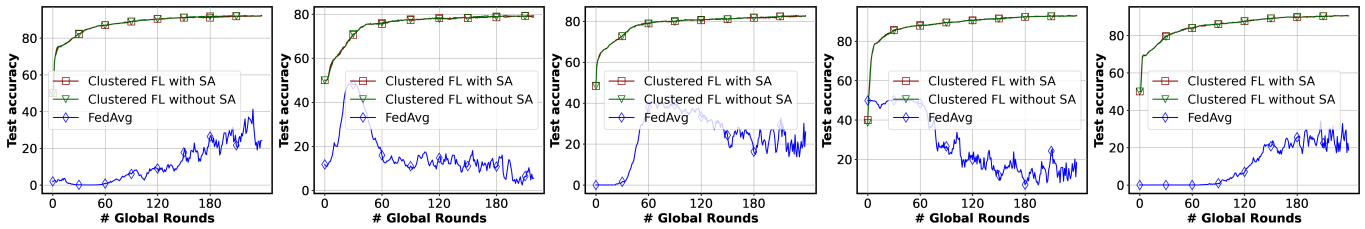
Setup. We consider clustered FL for image classification on the MNIST [37] and CIFAR-10 [38] datasets. Each user holds a local dataset sampled from one of K source distributions, where the data samples for each source distribution are realized from two distinct classes. Specifically, for both MNIST and CIFAR-10 datasets, which contain 10 classes, the data samples with labels $\{2j, 2j+1\}$ are distributed uniformly at random across the users $\{10j+1, \ldots, 10j+10\}$ for $j = 0, \ldots, 4$. Training is then performed using the CNN architectures from [1], where the number of model parameters is $d = 21840$ for MNIST and $d = 62006$ for CIFAR-10. The maximum number of adversarial and dropout users are $T = D = \lfloor\frac{N-3}{6}\rfloor$. For user dropouts, we consider the worst-case scenario for training, where maximum number of

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3366394



(a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4. (e) Cluster 5.

Fig. 2. Test accuracy vs number of iterations (MNIST dataset).



(a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4. (e) Cluster 5.

Fig. 3. Test accuracy vs number of iterations (CIFAR-10 dataset).

dropouts occur in the first round of online communication, i.e., $|\mathcal{U}_1| = |\mathcal{U}_2| = N - D$, as a result, the local gradients of the dropped users do not contribute to the global model at that iteration. At each iteration, D users drop out uniformly at random. The remaining hyperparameters are $L = 3$, $\eta = 0.001$, and $q = 2^{32} - 5$.

Model accuracy. We first evaluate the performance of our frameworks in terms of the test accuracy, with respect to the clustered FL benchmark (clustered FL without SA) from [23], which serves as our target accuracy, and the conventional (non-clustered) FL benchmark (*FedAvg*) from [1]. The accuracy of our frameworks is depicted as *clustered FL with SA*, as all our frameworks preserve the correctness of the secure computations, leading to the same final result. In Figs. 2 and 3, we report the average test accuracy of the users within each cluster for the MNIST and CIFAR-10 datasets, respectively. Our frameworks (clustered FL with SA), where the local gradients are aggregated in the finite field, achieve comparable test accuracy to the target benchmark (clustered FL without SA). In Fig. 2, we observe that clustered FL (with or without SA) achieves an average accuracy of 99% across all clusters, whereas *FedAvg* achieves 81.7% accuracy on average with a worst case accuracy of 48.1% (cluster 5). The performance improvement is even more significant for CIFAR-10 as observed in Fig. 3, where clustered FL (with or without SA) achieves an average accuracy of 87.8%, compared to the 46.91% accuracy of *FedAvg*.

Communication overhead. In Fig. 4(a), we compare the total online communication overhead (across all users) for the proposed frameworks with varying N while letting $K = 5$, and $L = 3$. We observe that CMGA significantly reduces the online communication overhead compared to CSGS, by up to $15.8\times$ since the intensive point-to-point communication overhead is transferred to the offline phase. The communication overhead is further reduced by SAMC by $4.12\times$ since the overhead of SAMC is $O(N(K + d))$, compared to the $O(NKd)$ overhead of CMGA . In Fig. 4(b), we further observe the impact of the multiplicative factor K on the per-user online communication overhead of CMGA, by letting $N = 200$, and varying K. We then set L accordingly to satisfy the recovery threshold $L = \frac{N-D-T}{K}$ for CSGS and CMGA, and $L = \frac{\frac{N-D-1}{2}-T+1}{K}$ for SAMC. As K increases, the communication overhead of CMGA increases linearly (as d is fixed), while having negligible impact on the communication overhead of SAMC (since $d >> K$). As such, SAMC reduces the per-user online communication overhead by up to $15.1\times$ compared to CMGA.

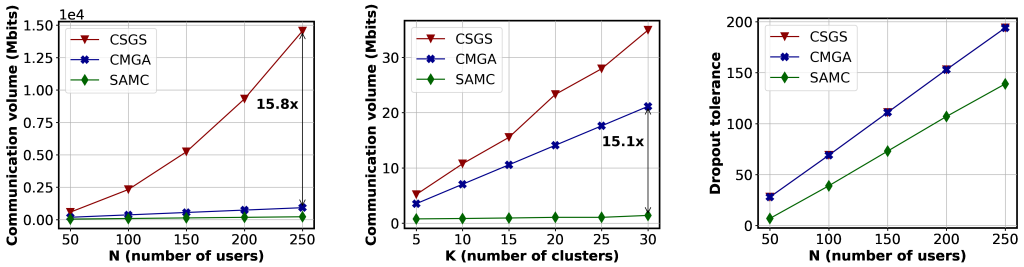
Dropout tolerance. In Fig. 4(c), we illustrate the maximum number of user dropouts that can be tolerated by each framework with varying number of users N, while keeping K and L fixed ($K = 5$, $L = 3$). We observe that CSGS and CMGA achieve a higher dropout tolerance compared to SAMC, which again reflects the trade-off between the dropout resilience and online communication overhead for the three frameworks.

Membership inference attacks. In Fig. 5, we demonstrate the impact of membership inference attacks [39] on the global model trained for each cluster (on CIFAR-10). Similar to [39–41], we consider a worst case scenario and perform the attack on the final model obtained for each cluster, as the attack performs better in the latter rounds of training (when models start to overfit to training data). Following [39], we report the attack performance in terms of precision, by measuring the fraction of samples interpreted as members that are actually members of the training dataset. In Fig. 5 we demonstrate the attack performance with varying number of users per cluster, with 500 data points per user, sampled from two distinct classes for each cluster. We observe that the attack performance degrades as the cluster size increases.

VII. Discussion

In this work, we focus on the honest-but-curious (passive) adversary model, as a first step for understanding more capa-

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3366394

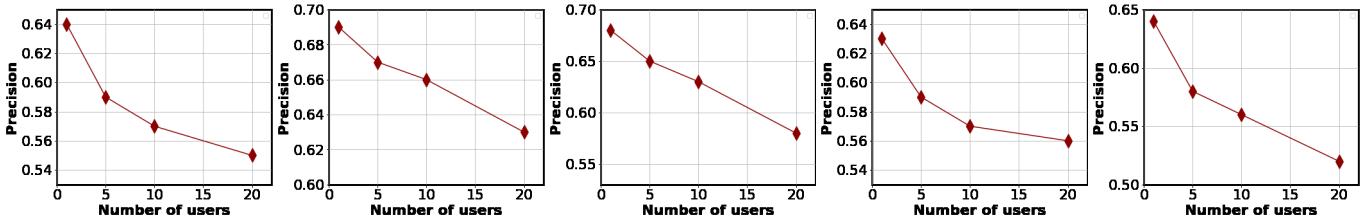


(a) Total online communication overhead with varying number of users (per training iteration).

(b) Per-user online communication overhead with varying number of clusters (per training iteration).

(c) Dropout tolerance for varying number of users.

Fig. 4. Performance comparisons for the three frameworks in terms of the communication overhead (with respect to the number of users and clusters) and dropout resilience (with respect to the number of users) on the MNIST dataset.



(a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4. (e) Cluster 5.

Fig. 5. Membership inference attack on the gradient aggregate of the clusters.

ble active (malicious) adversaries [42]. An interesting future direction is to extend our frameworks to the latter, who can modify the messages exchanged during protocol execution. One approach to achieve this is by leveraging Byzantine-resilient and verifiable secure multi-party computing mechanisms [42]. Verifiable secret sharing frameworks can ensure the correctness of the encoded messages sent from each user to the other parties [43]. The correctness of the polynomial computations sent from the users to the server, on the other hand, can be ensured by Reed-Solomon decoding, which can correctly identify the errors in the polynomial evaluations sent from the users to the server [44]. To ensure correct decoding in a network with up to A active adversaries, Reed-Solomon decoding requires two messages per error, hence the server needs $2A$ additional evaluations from the surviving users.

In addition to the encoding/decoding protocol, adversaries can also target the machine learning/training mechanism, by modifying their local datasets to inject unwanted behaviour into the global model [45]. Defending against such attacks requires secure outlier detection mechanisms as the local gradients are hidden during training (to preserve privacy), where local gradients from different users are compared without revealing their true value, and then outliers are removed during the aggregation of the local gradients at the server [42]. For clustered FL, doing so requires effective mechanisms for distinguishing the outliers that emerge from adversarial attacks from those that emerge from data heterogeneity.

VIII. Conclusions

In this work, we propose SA for clustered FL, to aggregate the local gradients for any cluster of users, without learning any information about the local gradients or cluster identities of the users. Our framework can achieve linear communication complexity, while ensuring formal information-theoretic privacy guarantees. Future directions include extending our mechanisms to active (malicious) adversaries who can modify the messages or datasets adversarially [42], and integrating our frameworks with authenticated key agreement mechanisms to ensure the integrity and authenticity of the messages exchanged against malicious adversaries. Another interesting direction is to further enhance the computational efficiency by offloading the computational overhead of polynomial interpolations to the offline phase, by leveraging trusted execution environments or crypto-service providers [46].

Appendix A
Finite Field Representations

The local gradient $\mathbf{g}_i$ of user i is represented in the finite field $\mathbb{F}_q$ as $\overline{\mathbf{g}}_i \triangleq \rho(\mathbf{g}_i) \mod q$, using a stochastic quantization function [42, 47],

$$\rho(x) \triangleq \begin{cases} \lfloor lx \rfloor & \text{with probability} \quad 1-(lx-\lfloor lx \rfloor) \\ \lfloor lx \rfloor + 1 & \text{with probability} \quad lx - \lfloor lx \rfloor \end{cases} \tag{41}$$

operating element-wise, where l controls the quantization loss (set to 2^{20} in the experiments), and the modulo operation maps the negative integers in the second half of the finite field. Prime q is selected large enough to avoid a wrap-around which may cause overflow errors. After recovering the aggregate of the gradients for each cluster, the server updates the models,

$$\mathbf{w}_k \leftarrow \mathbf{w}_k - \frac{\eta}{Nl}\rho^{-1}\Big(\sum_{i\in\mathcal{S}_k\cap\mathcal{U}} \overline{\mathbf{g}}_i\Big) \quad \forall k \in [K] \tag{42}$$

where $\rho^{-1}:\mathbb{F}_q\to\mathbb{R}$ is a demapping function that converts the gradients back to the real domain,

$$\rho^{-1}(\bar{x})=\begin{cases}\bar{x} & \text{if } 0\le\bar{x}<\frac{q-1}{2}\\ \bar{x}-q & \text{if } \frac{q-1}{2}\le\bar{x}\le q\end{cases} \tag{43}$$

Appendix B
Information-theoretic Privacy

We now demonstrate the information-theoretic privacy of each framework against any set $\mathcal{T}$ of $|\mathcal{T}|=T$ adversarial users (the proof for any $|\mathcal{T}|<T$ follows the same steps). The set of honest users is denoted by $\mathcal{H}=[N]\backslash\mathcal{T}$. Without loss of generality, we let $\mathcal{T}=[T]$, as the same analysis holds for any set $\mathcal{T}\subset[N]$ of size T. For the analysis, we consider the worst-case scenario where all messages are communicated across the users, i.e., users declared as dropped are only delayed, and their messages are eventually received by the adversaries [9]. The number of surviving users at the final communication round, denoted by $|\mathcal{U}_2|$, is assumed to be equal to the recovery threshold of the corresponding framework (the same analysis also holds for a larger number of surviving users). We next demonstrate the privacy analysis for each framework.

CSGS. For this framework, the mutual information condition in (40) can be written as,

$$\begin{aligned}
&I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}|\\
&\quad\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]})\\
&=I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]},\{\alpha_i\}_{i\in[N]})+I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\\
&\{\alpha_i\}_{i\in[N]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]})
\end{aligned}\tag{44}$$

$$\begin{aligned}
&=I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\quad\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]},\{\alpha_i\}_{i\in[N]})+0
\end{aligned}\tag{45}$$

$$\begin{aligned}
&=H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\\
&\quad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})-H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})
\end{aligned}\tag{46}$$

where (44) follows from the chain rule of mutual information, and (45) holds since the public parameters $\{\alpha_i\}_{i\in[N]}$ are generated independently from the locally generated gradients and random masks, (46) follows from the chain rule of entropy and the fact that there is no uncertainty in $\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{T}\\ j\in\mathcal{T}}}$ given $\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}}$, $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$. For the second term in (46), we have,

$$\begin{aligned}
&H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\\
&\quad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})\\
&=H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\\
&\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\quad\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})
\end{aligned}\tag{47}$$

$$\begin{aligned}
&=H(\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\quad\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]})\\
&+H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in[T]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\\
&\quad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})
\end{aligned}\tag{48}$$

$$\begin{aligned}
&=H(\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]})+H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in[T]}})
\end{aligned}\tag{49}$$

$$=0+H(\{\mathbf{v}_{il}\}_{i\in\mathcal{H},l\in[T]}) \tag{50}$$

$$=(N-T)T\frac{d}{L}\log q \tag{51}$$

where (47) is from the chain rule. By denoting $\mathcal{U}_2=\{U_1,\dots,U_{KL+T}\}$,

$$\begin{aligned}
\big[\widetilde{\mathbf{g}}_{U_1}\ \ \cdots\ \ \widetilde{\mathbf{g}}_{U_{KL+T}}\big]=\Big[&\sum_{i\in\mathcal{S}_1\cap\mathcal{U}_1}\overline{\mathbf{g}}_{i1}\ \ \sum_{i\in\mathcal{S}_1\cap\mathcal{U}_1}\overline{\mathbf{g}}_{i2}\ \cdots\\
&\sum_{i\in\mathcal{S}_K\cap\mathcal{U}_1}\overline{\mathbf{g}}_{iL}\ \ \sum_{i\in\mathcal{U}_1}\mathbf{v}_{i1}\ \cdots\ \sum_{i\in\mathcal{U}_1}\mathbf{v}_{iT}\Big]\mathbf{Q}
\end{aligned}\tag{52}$$

where $\mathbf{Q}$ is a $(KL+T)\times(KL+T)$ MDS (maximum distance separable) matrix,

$$\mathbf{Q}\triangleq\begin{bmatrix}1 & \alpha_{U_1} & \cdots & \alpha_{U_1}^{KL-1} & \alpha_{U_1}^{KL} & \cdots & \alpha_{U_1}^{KL+T-1}\\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 1 & \alpha_{U_{KL+T}} & \cdots & \alpha_{U_{KL+T}}^{KL-1} & \alpha_{U_{KL+T}}^{KL} & \cdots & \alpha_{U_{KL+T}}^{KL+T-1}\end{bmatrix}^{\mathrm{T}}$$

hence $\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2}$ is invertible to $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{il}\}_{k\in[K],l\in[L]}$, $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$, from which (48) follows. Equation (49) holds since $\{\mathbf{v}_{il}\}_{i\in\mathcal{H},l\in[T]}$ is generated uniformly at random from $\mathbb{F}_q$ (and independently from other elements). Note that given $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$ and $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, one can compute $\{\widetilde{\mathbf{g}}_{ij}\}_{i\in\mathcal{T},j\in\mathcal{T}}$ from (10). Since $\mathcal{T}=[T]$, one can find from (11) and (12) that,

$$\begin{aligned}
&\begin{bmatrix}\sum_{i\in\mathcal{U}_1}\widetilde{\mathbf{g}}_{i1}^{\mathrm{T}}-\sum_{k\in[K]}\sum_{l\in[L]}\alpha_1^{(k-1)L+l-1}\Big(\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{il}^{\mathrm{T}}\Big)\\ \vdots\\ \sum_{i\in\mathcal{U}_1}\widetilde{\mathbf{g}}_{iT}^{\mathrm{T}}-\sum_{k\in[K]}\sum_{l\in[L]}\alpha_T^{(k-1)L+l-1}\Big(\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_{il}^{\mathrm{T}}\Big)\end{bmatrix}^{\mathrm{T}}\\
&=\Big[\sum_{i\in\mathcal{U}_1}\mathbf{v}_{i1}\ \cdots\ \sum_{i\in\mathcal{U}_1}\mathbf{v}_{iT}\Big]\mathbf{A},
\end{aligned}\tag{53}$$

where

$$\mathbf{A}\triangleq\begin{bmatrix}\alpha_1^{KL} & \cdots & \alpha_T^{KL}\\ \vdots & \ddots & \vdots\\ \alpha_1^{KL+T-1} & \cdots & \alpha_T^{KL+T-1}\end{bmatrix} \tag{54}$$

is a $T\times T$ MDS matrix (hence invertible). Therefore, there is no uncertainty in $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$ given

$\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, $\{\overline{\mathbf{g}}_i, c_i\}_{i\in\mathcal{T}}$, $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, $\{\widetilde{\mathbf{g}}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, $\{\alpha_i\}_{i\in[N]}$ from which (50) holds. Equation (51) holds as the entropy of a uniform random variable over the alphabet $\mathcal{B}$ is $\log|\mathcal{B}|$ [48]. Next, the first term in (46) can be bounded as:

$$\begin{aligned}
&H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\\
&\qquad\qquad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})\\
&=H(\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\\
&\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\\
&\qquad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}) \qquad (55)\\
&=H(\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\\
&\{\alpha_i\}_{i\in[N]})+H(\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\\
&\qquad\qquad\{\alpha_i\}_{i\in[N]}) \qquad (56)\\
&\leq H(\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}) \qquad (57)\\
&\leq (N-T)T\frac{d}{L}\log q \qquad (58)
\end{aligned}$$

where (55) is from the chain rule of entropy; (56) follows from (52) as $\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2}$ is invertible to $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$; (57) holds since there is no uncertainty in $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$ given $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$, $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, $\{\widetilde{\mathbf{g}}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, $\{\alpha_i\}_{i\in[N]}$ from (53), and that conditioning cannot increase entropy; (58) holds since uniform distribution maximizes entropy. By combining (46), (51), (58) with the non-negativity of mutual information,

$$\begin{aligned}
&0\leq I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\widetilde{\mathbf{g}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{g}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}) \qquad (59)\\
&\leq (N-T)T\frac{d}{L}\log q-(N-T)T\frac{d}{L}\log q \qquad (60)\\
&=0 \qquad (61)
\end{aligned}$$

which completes the proof.

CMGA. For this framework, the mutual information from (40) can be written as,

$$\begin{aligned}
&I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}})\\
&=I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in[N]\\ j\in\mathcal{T}}}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\\
&\{\alpha_i\}_{i\in[N]})+I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\alpha_i\}_{i\in[N]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}}) \qquad (62)\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})\\
&-H(\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]}) \qquad (63)
\end{aligned}$$

where (63) follows from the chain rule, along with the fact that there is no uncertainty in $\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{T}\\ j\in\mathcal{T}}}$ given $\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}}$, $\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}}$, and that the public parameters $\{\alpha_i\}_{i\in[N]}$ are generated independently from the locally generated gradients and random masks. For the second term in (63), we have,

$$\begin{aligned}
&H(\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})\\
&=H(\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]}) \qquad (64)\\
&=H(\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in[N]\\ k\in[k]}},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in[N]\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\\
&\qquad\{\alpha_i\}_{i\in[N]})+H(\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) \qquad (65)\\
&=H(\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]},\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\{\mathbf{r}_{ik}\}_{\substack{i\in[N]\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]})\\
&+H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in[T]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\\
&\{\mathbf{r}_{ik}\}_{\substack{i\in[N]\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) \qquad (66)\\
&=0+H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in[T]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in[N]},\\
&\{\mathbf{r}_{ik}\}_{\substack{i\in[N]\\ k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) \qquad (67)\\
&=H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in[T]}})+H(\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) \qquad (68)\\
&=(N-T)T\frac{d}{L}\log q+(N-T)Kd\log q \qquad (69)
\end{aligned}$$

where (64) holds since given $\overline{\mathbf{g}}_i$, c_i, the uncertainty in $\mathbf{x}_{ik}$ is due to $\mathbf{r}_{ik}$ for all $k\in[K]$; (65) is from the chain rule of entropy and that the random vectors $\{\mathbf{r}_{ik}\}_{i\in\mathcal{H},k\in[K]}$ are generated independently; (66) holds since similar to (52), $\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2}$ is invertible to $\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]}$, $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$ and given $\{\mathbf{r}_{ik}\}_{i\in[N],k\in[K]}$, $\{\widetilde{\mathbf{r}}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$ is invertible to $\{\mathbf{v}_{il}\}_{i\in\mathcal{H},l\in[T]}$. Note that given $\{\mathbf{r}_{ik}\}_{i\in\mathcal{T},k\in[K]}$ and $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, one can compute $\{\widetilde{\mathbf{r}}_{ij}\}_{i\in\mathcal{T},j\in\mathcal{T}}$ from (16), and given $\{\mathbf{r}_{ik}\}_{i\in[N],k\in[K]}$, one can compute $\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]}$. Then, by using (18) and

(19), one can find that,

$$\begin{bmatrix} \sum_{i\in\mathcal{U}_1}\widetilde{\mathbf{r}}_{i1}^{\mathrm{T}} - \sum_{k\in[K]}\sum_{l\in[L]}\alpha_1^{(k-1)L+l-1}\Big(\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ikl}^{\mathrm{T}}\Big) \\ \vdots \\ \sum_{i\in\mathcal{U}_1}\widetilde{\mathbf{r}}_{iT}^{\mathrm{T}} - \sum_{k\in[K]}\sum_{l\in[L]}\alpha_T^{(k-1)L+l-1}\Big(\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ikl}^{\mathrm{T}}\Big) \end{bmatrix}^{\mathrm{T}} = \Big[\sum_{i\in\mathcal{U}_1}\mathbf{v}_{i1}\ \cdots\ \sum_{i\in\mathcal{U}_1}\mathbf{v}_{iT}\Big]\mathbf{A}, \tag{70}$$

where $\mathbf{A}$ is the $T\times T$ MDS matrix (invertible) from (54), hence there is no uncertainty in $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$ given $\{\mathbf{r}_{ik}\}_{i\in[N],k\in[k]}$, $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, $\{\widetilde{\mathbf{r}}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, $\{\alpha_i\}_{i\in[N]}$, from which (67) holds. Equation (68) is from the independence of generated randomness, (69) is from the entropy of uniform randomness. The first term in (63) can be bounded as,

$$\begin{aligned}
&H(\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]})\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\qquad\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]}) &(71)\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\alpha_i\}_{i\in[N]})\\
&+H(\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\qquad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}) &(72)\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\\
&\qquad\{\alpha_i\}_{i\in[N]}) &(73)\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\alpha_i\}_{i\in[N]})\\
&+H(\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\qquad\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}) &(74)\\
&=H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\\
&\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\alpha_i\}_{i\in[N]})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\\
&\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}},\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\alpha_i\}_{i\in[N]})+0 &(75)\\
&\le H(\{\mathbf{x}_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}})+H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) &(76)\\
&\le (N-T)Kd\log q+(N-T)T\frac{d}{L}\log q &(77)
\end{aligned}$$

where (71) holds since given $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{i\in\mathcal{T},k\in[k]}$, there is no uncertainty in $\{\mathbf{x}_{ik}\}_{i\in\mathcal{T},k\in[K]}$; (72) is from the chain rule; and (73) holds as $\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2}$ is invertible to $\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]}$, and $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$. Note that given $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$, $\{\mathbf{r}_{ik}\}_{i\in\mathcal{T},k\in[k]}$, one can compute $\{\mathbf{x}_{ik}\}_{i\in\mathcal{T},k\in[K]}$ from (17), and given $\{\mathbf{x}_{ik}\}_{i\in[N],k\in[K]}$, one can compute $\{\sum_{i\in\mathcal{U}_1}\mathbf{x}_{ik}\}_{k\in[K]}$. Finally, $\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]}$ can be computed from $\{\sum_{i\in\mathcal{U}_1}\mathbf{x}_{ik}\}_{k\in[K]}$ and $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$ by using (21). Therefore, there is no uncertainty in $\{\sum_{i\in\mathcal{U}_1}\mathbf{r}_{ik}\}_{k\in[K]}$ given $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$, $\{\mathbf{r}_{ik}\}_{i\in\mathcal{T},k\in[k]}$, $\{\mathbf{x}_{ik}\}_{i\in\mathcal{H},k\in[K]}$, from which (74) holds. Similarly, as can be observed from (70), there is no uncertainty in $\{\sum_{i\in\mathcal{U}_1}\mathbf{v}_{il}\}_{l\in[T]}$ given $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, $\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}}$, $\{\mathbf{r}_{ik}\}_{i\in\mathcal{T},k\in[k]}$, $\{\mathbf{v}_{il}\}_{i\in\mathcal{T},l\in[T]}$, $\{\mathbf{x}_{ik}\}_{i\in\mathcal{H},k\in[K]}$, $\{\widetilde{\mathbf{r}}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, $\{\alpha_i\}_{i\in[N]}$, from which (75) holds; (76) holds as conditioning cannot increase entropy; (77) holds as uniform distribution maximizes entropy. By combining (63), (69), and (77) with the non-negativity of mutual information, we have that,

$$\begin{aligned}
0\le& I(\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{H}};\{\mathbf{x}_{ik}\}_{\substack{i\in[N]\\k\in[K]}},\{\widetilde{\mathbf{r}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\{\alpha_i\}_{i\in[N]}|\\
&\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,c_i\}_{i\in\mathcal{T}},\{\mathbf{r}_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[k]}},\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{T}\\l\in[T]}})\\
\le&(N-T)Kd\log q+(N-T)T\frac{d}{L}\log q\\
&-(N-T)Kd\log q-(N-T)T\frac{d}{L}\log q &(78)\\
=&0 &(79)
\end{aligned}$$

which completes the proof.

SAMC. In the following, we let $C\triangleq 2(KL+T-1)+1$ and,

$$\begin{aligned}
\mathcal{A}\triangleq\big\{&\{\alpha_i\}_{i\in[N]},\{\beta_m\}_{m\in[KL+T]},\\
&\{\theta_m\}_{m\in\{KL+1,\ldots,2(KL+T-1)+1\}},\{\lambda_m\}_{m\in\{N-T\}}\big\}
\end{aligned}\tag{80}$$

Then, the mutual information from (40) can be written as:

$$\begin{aligned}
&I(\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}};\{\mathbf{x}_i\}_{i\in[N]},\{y_{ik}\}_{\substack{i\in[N]\\k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2},\\
&\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\mathcal{A}|\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\\
&\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},\\
&\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}})\\
=&H(\{\mathbf{x}_i\}_{i\in[N]},\{y_{ik}\}_{\substack{i\in[N]\\k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\\
&\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\\
&\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}},\mathcal{A})
\end{aligned}$$

$$- H(\{\mathbf{x}_i\}_{i\in[N]}, \{y_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},$$
$$\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{r}_i, z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A}) \quad (81)$$

For the second term in (81), we find that:

$$H(\{\mathbf{x}_i\}_{i\in[N]}, \{y_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},$$
$$\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{r}_i, z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A})$$
$$= H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}, \{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}, \{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},$$
$$\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{r}_i, z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A}) \quad (82)$$
$$= H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} |$$
$$\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \mathcal{A})$$
$$+ H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]},$$
$$\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}, \mathcal{A}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},$$
$$\{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in\mathcal{T}}, \{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}, \mathcal{A}) \quad (83)$$
$$= H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} |$$
$$\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \mathcal{A})$$
$$+ H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (84)$$
$$= H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} |$$
$$\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \mathcal{A})$$
$$+ H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (85)$$
$$= H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},$$
$$\{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (86)$$
$$= H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},$$
$$\{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (87)$$
$$= H(\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]},$$
$$\{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \mathcal{A}) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (88)$$
$$= H(\{\sum_{j\in\mathcal{U}_1}\phi_j(\alpha_i)\psi_j(\alpha_i) - \widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} |$$
$$\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]}, \{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},$$
$$\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \mathcal{A}) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A})$$
$$+ H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (89)$$
$$= H(\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}, \{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}, \{\overline{\mathbf{g}}_i, b_{ik}\}_{\substack{i\in[N]\\k\in[K]}},$$
$$\{\mathbf{v}_{il}, u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,KL+T\}}}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \{\mathbf{r}_i\}_{i\in[N]},$$
$$\{z_{ik}\}_{\substack{i\in[N]\\k\in[K]}}, \{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}, \mathcal{A}) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (90)$$
$$= H(\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}, \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A})$$
$$+ H(\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2} | \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A}) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (91)$$
$$= 0 + H(\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2} | \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A})$$
$$+ H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}})$$
$$+ H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (92)$$
$$= H(\{\mathbf{n}_l\}_{l\in\{KL+1,\ldots,C\}} | \{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A})$$
$$+ H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}})$$
$$+ H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (93)$$
$$= H(\{[\sum_{j\in[N]}\lambda_1^{j-1}\mathbf{n}_{jl}^{\mathrm{T}} \ \cdots \ \sum_{j\in[N]}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}^{\mathrm{T}}]^{\mathrm{T}}\}_{l\in\{KL+1,\ldots,C\}} |$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\ldots,C\}}}, \mathcal{A}) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A})$$
$$+ H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}} | \mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \quad (94)$$

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3366394

$$= H\big(\big\{\big[\sum_{j\in\mathcal{H}}\lambda_1^{j-1}\mathbf{n}_{jl}^{\mathrm{T}}\ \cdots\ \sum_{j\in\mathcal{H}}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}^{\mathrm{T}}\big]^{\mathrm{T}}\big\}_{l\in\{KL+1,\ldots,C\}}|\ \mathcal{A}\big) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\mathcal{A}) + H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \tag{95}$$

$$= H\big(\big\{\big[\mathbf{n}_{(T+1)l}\ \cdots\ \mathbf{n}_{Nl}\big]\mathbf{B}\big\}_{l\in\{KL+1,\ldots,C\}}|\mathcal{A}\big) + H(\{e_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\mathcal{A}) + H(\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\mathcal{A}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \tag{96}$$

$$= H(\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in\{KL+1,\ldots,C\}}}) + H(\{u_{il}\}_{\substack{i\in\mathcal{H}\\ l\in\{KL+1,\ldots,KL+T\}}}) + H(\{\mathbf{v}_{il}\}_{\substack{i\in\mathcal{H}\\ l\in\{KL+1,\ldots,KL+T\}}}) + H(\{z_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) + H(\{\mathbf{r}_i\}_{i\in\mathcal{H}}) \tag{97}$$

$$= (C-KL)(N-T)\frac{d}{L(N-T)}\log q + (N-T)T\log q + (N-T)T\frac{d}{L}\log q + (N-T)K\log q + (N-T)d\log q \tag{98}$$

where (82) holds since given $\overline{\mathbf{g}}_i$ and $\{b_{ik}\}_{k\in[K]}$, the uncertainty in $\mathbf{x}_i$ and $\{y_{ik}\}_{k\in[K]}$ is due to the uncertainty in $\mathbf{r}_i$ and $\{z_{ik}\}_{k\in[K]}$; (83) follows from chain rule of entropy; (84) holds since $\{z_{ik}\}_{i\in\mathcal{H},k\in[K]}$ and $\{\mathbf{r}_i\}_{i\in\mathcal{H}}$ are independent uniformly random vectors in $\mathbb{F}_q$. In (85), we define:

$$\mathbf{m}_{ij} \triangleq \sum_{l=KL+1}^{KL+T}\mathbf{v}_{il}\prod_{m\in[KL+T]\setminus\{l\}}\frac{\alpha_j-\beta_m}{\beta_l-\beta_m} \quad \forall i\in\mathcal{H}, j\in\mathcal{T} \tag{99}$$

and (86) is from the chain rule and independence of generated randomness. In (87), we define:

$$e_{ij} \triangleq \sum_{l=KL+1}^{KL+T}u_{il}\prod_{m\in[KL+T]\setminus\{l\}}\frac{\alpha_j-\beta_m}{\beta_l-\beta_m} \tag{100}$$

whereas (88) follows from the chain rule and independence of generated randomness. We next let $\gamma_{jl} := \prod_{m\in[KL+T]\setminus\{l\}}\frac{\alpha_j-\beta_m}{\beta_l-\beta_m}$ denote the Lagrange coefficients in (100)-(99). Then,

$$\big[\mathbf{m}_{i1}\ \cdots\ \mathbf{m}_{iT}\big] = \big[\mathbf{v}_{i,KL+1}\ \cdots\ \mathbf{v}_{i,KL+T}\big]\mathbf{M}, \tag{101}$$

$$\big[e_{i1}\ \cdots\ e_{iT}\big] = \big[u_{i,KL+1}\ \cdots\ u_{i,KL+T}\big]\mathbf{M}, \tag{102}$$

where

$$\mathbf{M} \triangleq \begin{bmatrix}\gamma_{1,KL+1} & \cdots & \gamma_{T,KL+1}\\ \vdots & \ddots & \vdots\\ \gamma_{1,KL+T} & \cdots & \gamma_{T,KL+T}\end{bmatrix}, \tag{103}$$

is a $T\times T$ MDS matrix (hence invertible) from the MDS property of Lagrange coefficients [49]. Hence, one can recover $\{\mathbf{v}_{il},u_{il}\}_{i\in\mathcal{H},l\in\{KL+1,\ldots,KL+T\}}$ given $\{\mathbf{m}_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, $\{e_{ij}\}_{i\in\mathcal{H},j\in\mathcal{T}}$, and $\mathcal{A}$. Then, (90) holds as there is no uncertainty in $\{\sum_{j\in\mathcal{U}_1}\phi_j(\alpha_i)\psi_j(\alpha_i)\}_{i\in\mathcal{U}_2}$ given $\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in[N]\\ k\in[K]}}$, $\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,KL+T\}}}$, $\{\mathbf{m}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}$, $\{e_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}$, and $\mathcal{A}$ since $\phi_j(\alpha_i)$, $\psi_j(\alpha_i)$ can be computed from (37) and (38) given $\overline{\mathbf{g}}_j$, $\{\mathbf{v}_{jl}\}_{l\in\{KL+1,\ldots,KL+T\}}$ and b_{jk},$\{u_{jl}\}_{l\in\{KL+1,\ldots,KL+T\}}$, respectively. Equation (92) holds since $\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}$ correspond to evaluation points of the degree of $C-1$ polynomial $v(\alpha)$ from (30). As any polynomial of degree $C-1$ can be uniquely interpolated from any set of C evaluation points, $v(\alpha)$ can be reconstructed from $|\mathcal{U}_2|=C$ evaluations $\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}$, from which one can recover $\mathbf{n}_l=v(\theta_l)$ for all $l\in\{KL+1,\ldots,C\}$. Then, given $\{\mathbf{n}_l\}_{l\in\{KL+1,\ldots,C\}}$, $\{\mathbf{n}_{jl}\}_{j\in\mathcal{T},l\in\{KL+1,\ldots,C\}}$, and $\mathcal{A}$, one can reconstruct $\{\mathbf{n}_{jl}\}_{j\in\mathcal{H},l\in\{KL+1,\ldots,C\}}$ as,

$$\begin{aligned}&\big[\mathbf{n}_{(T+1)l}\ \cdots\ \mathbf{n}_{Nl}\big]\mathbf{B}\\ &= \mathbf{n}_l - \Big[\textstyle\sum_{j\in[T]}\lambda_1^{j-1}\mathbf{n}_{jl}\ \cdots\ \sum_{j\in[T]}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}\Big]\\ &= \Big[\textstyle\sum_{j\in[N]}\lambda_1^{j-1}\mathbf{n}_{jl}\ \cdots\ \sum_{j\in[N]}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}\Big]\\ &- \Big[\textstyle\sum_{j\in[T]}\lambda_1^{j-1}\mathbf{n}_{jl}\ \cdots\ \sum_{j\in[T]}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}\Big]\\ &= \Big[\textstyle\sum_{j=T+1}^{N}\lambda_1^{j-1}\mathbf{n}_{jl}\ \cdots\ \sum_{j=T+1}^{N}\lambda_{N-T}^{j-1}\mathbf{n}_{jl}\Big]\end{aligned}$$

for all $l\in\{KL+1,\ldots,C\}$, where

$$\mathbf{B} \triangleq \begin{bmatrix}\lambda_1^T & \cdots & \lambda_{N-T}^T\\ \vdots & \ddots & \vdots\\ \lambda_1^{N-1} & \cdots & \lambda_{N-T}^{N-1}\end{bmatrix} \tag{104}$$

is an $(N-T)\times(N-T)$ MDS matrix (invertible). Equation (93) holds since polynomial $v(\alpha)$ has degree $C-1$, which can be uniquely constructed from any set of C evaluation points. Hence, there is a bijective mapping between any C interpolation points, $\{v(\theta_l)\}_{l\in[C]}$, where $v(\theta_l)=0$ for $l\in[KL]$, and $v(\theta_l)=\mathbf{n}_l$ for $l\in\{KL+1,\ldots,C\}$, and the set of local computations $\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{U}_2}$ where $|\mathcal{U}_2|=C$. Equation (97) holds from (101) and (102). Finally, (98) is from the entropy of uniform random variables. Next, the first term in (81) can be bounded as follows:

$$\begin{aligned}&H(\{\mathbf{x}_i\}_{i\in[N]},\{y_{ik}\}_{\substack{i\in[N]\\ k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\\ &\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\\ &\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,C\}}},\mathcal{A})\\ &= H(\{\mathbf{x}_i\}_{i\in\mathcal{H}},\{y_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\\ &\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\\ &\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,C\}}},\mathcal{A})\\ &\le H(\{\mathbf{x}_i\}_{i\in\mathcal{H}}) + H(\{y_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) + H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}})\\ &+ H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}|\\ &\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\\ &\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\ l\in\{KL+1,\ldots,C\}}},\{\mathbf{x}_i\}_{i\in\mathcal{H}},\\ &\{y_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}},\mathcal{A})\\ &= H(\{\mathbf{x}_i\}_{i\in\mathcal{H}}) + H(\{y_{ik}\}_{\substack{i\in\mathcal{H}\\ k\in[K]}}) + H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}})\\ &+ H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\ j\in\mathcal{T}}}) + H(\{f(\theta_l)\}_{l\in[C-T]},\\ &\{f(\alpha_i)\}_{i\in[T]}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\ k\in[K]}},\{\mathbf{r}_i\}_{i\in\mathcal{T}},\end{aligned} \tag{105}$$

$$\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,C\}}},$$
$$\{\mathbf{x}_i\}_{i\in\mathcal{H}},\{y_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\mathcal{A}) \tag{107}$$

$$= H(\{\mathbf{x}_i\}_{i\in\mathcal{H}}) + H(\{y_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}})$$
$$+ H(\{f(\theta_l)\}_{l\in\{KL+1,\dots,C-T\}}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},$$
$$\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{r}_i\}_{i\in\mathcal{T}},\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},$$
$$\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,KL+T\}}},\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,C\}}},\{\mathbf{x}_i\}_{i\in\mathcal{H}},$$
$$\{y_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}},\mathcal{A}) \tag{108}$$

$$\leq H(\{\mathbf{x}_i\}_{i\in\mathcal{H}}) + H(\{y_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}) + H(\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) + H(\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}) + H(\{f(\theta_l)\}_{l\in\{KL+1,\dots,C-T\}}) \tag{109}$$

$$\leq (N-T)d\log q + (N-T)K\log q + (N-T)T\log q + (N-T)T\frac{d}{L}\log q + T(N-T)\frac{d}{L(N-T)}\log q + (C-T-KL)\frac{d}{L}\log q \tag{110}$$

where (106) is from the chain rule and that conditioning cannot increase entropy. Equation (107) holds since $\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}$ correspond to C evaluations of a degree $C-1$ polynomial, $f(\alpha)$ from (36) for $\alpha\in\{\alpha_i\}_{i\in\mathcal{U}_2}$. Since a polynomial of degree $C-1$ can be uniquely reconstructed from any set of C evaluation points, there is a bijective mapping between the C interpolation points $\{f(\theta_l)\}_{l\in[C-T]},\{f(\alpha_i)\}_{i\in[T]}$ and the C local computations $\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2}$. Note that there is no uncertainty in $\{f(\theta_l)\}_{l\in[KL]}$ given $\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]}$, and $\mathcal{A}$, which follows from (39) and that $\theta_l=\beta_l$ for $l\in[KL]$. Next, note that one can compute $\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{T}\\j\in\mathcal{T}}}$, $\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{T}\\j\in\mathcal{T}}}$, and $\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{T}\\j\in\mathcal{T}}}$ given $\{\mathbf{r}_i\}_{i\in\mathcal{T}}$, $\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$, $\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,KL+T\}}}$, $\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,C\}}}$, and $\mathcal{A}$ from (24), (26), and (28). Moreover, given $\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}}$, one can compute $\{\widetilde{\mathbf{n}}_i\}_{i\in\mathcal{T}}$ from (29), and given $\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$, $\{\mathbf{r}_i\}_{i\in\mathcal{T}}$, $\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$, one can compute $\{\mathbf{x}_i\}_{i\in\mathcal{T}}$, $\{y_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$ from (32) and (33). Therefore, there is no uncertainty in $\{f(\alpha_i)\}_{i\in[T]}=\{\widetilde{\mathbf{a}}_i\}_{i\in[T]}$ given $\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$, $\{\mathbf{r}_i\}_{i\in\mathcal{T}}$, $\{z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}}$, $\{\mathbf{x}_i\}_{i\in\mathcal{H}}$, $\{y_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}}$, $\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,KL+T\}}}$, $\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,C\}}}$, $\{\widetilde{z}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}$, $\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}$, $\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in\mathcal{H}\\j\in\mathcal{T}}}$, and $\mathcal{A}$ according to (35), from which (108) holds. Finally, (110) holds since uniform distribution maximizes entropy. By combining (98), (110), and (81) with the non-negativity of mutual information,

$$0 \leq I(\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{H}\\k\in[K]}};\{\mathbf{x}_i\}_{i\in[N]},\{y_{ik}\}_{\substack{i\in[N]\\k\in[K]}},\{\widetilde{z}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},$$
$$\{\widetilde{\mathbf{a}}_i\}_{i\in\mathcal{U}_2},\{\widetilde{\mathbf{r}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\{\widetilde{\mathbf{n}}_{ij}\}_{\substack{i\in[N]\\j\in\mathcal{T}}},\mathcal{A}|\{\sum_{i\in\mathcal{S}_k\cap\mathcal{U}_1}\overline{\mathbf{g}}_i\}_{k\in[K]},$$
$$\{\overline{\mathbf{g}}_i,b_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{r}_i,z_{ik}\}_{\substack{i\in\mathcal{T}\\k\in[K]}},\{\mathbf{v}_{il},u_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,KL+T\}}},$$
$$\{\mathbf{n}_{il}\}_{\substack{i\in\mathcal{T}\\l\in\{KL+1,\dots,C\}}})$$
$$\leq (C-KL)\frac{d}{L}\log q + (N-T)T\log q + (N-T)T\frac{d}{L}\log q + K(N-T)\log q + (N-T)d\log q - (N-T)d\log q - (N-T)K\log q - (N-T)T\log q - (N-T)T\frac{d}{L}\log q - (C-KL)\frac{d}{L}\log q \tag{111}$$
$$= 0 \tag{112}$$

which completes the proof.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Int. Conf. on Artificial Int. and Stat. (AISTATS)*, 2017, pp. 1273–1282.

[2] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM Conf. on Computer and Communications Security (CCS)*, 2015.

[3] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 14 747–14 756.

[4] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *ACM Conf. on Comp. and Comm. Security (CCS)*, 2017.

[6] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova, "Secure single-server aggregation with (poly) logarithmic overhead," in *ACM Conf. on Computer and Communications Security (CCS)*, 2020.

[7] J. So, B. Güler, and A. S. Avestimehr, "Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning," *IEEE Journal on Selected Areas in Information Theory*, 2021.

[8] Y. Zhao and H. Sun, "Information theoretic secure aggregation with user dropouts," *IEEE Trans. Inf. Th.*, 2022.

[9] J. So, C.-S. Yang, S. Li, Q. Yu, R. E Ali, B. Guler, and S. Avestimehr, "Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning," *Proc. of Machine Learning and Systems (MLSys)*, 2022.

[10] T. Jahani-Nezhad, M. A. Maddah-Ali, S. Li, and G. Caire, "Swiftagg+: Achieving asymptotically optimal communication load in secure aggregation for federated learning," *IEEE Journal on Sel. Areas in Comm.*, 2023.

[11] A. R. Elkordy and A. S. Avestimehr, "Heterosag: Secure aggregation with heterogeneous quantization in federated learning," *IEEE Transactions on Communications*, vol. 70, no. 4, pp. 2372–2386, 2022.

[12] K. Wan, H. Sun, M. Ji, and G. Caire, "Information theoretic secure aggregation with uncoded groupwise keys," *arXiv:2204.11364v2*, 2022.

[13] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," *J. Priv. Confidentiality*, vol. 7, no. 3, pp. 17–51, 2016.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM Conf. on Computer and Communications Security (CCS)*, 2016.

[15] P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete gaussian mechanism for federated learning with secure aggregation," in *International Conference on Machine Learning (ICML)*, 2021, pp. 5201–5212.

[16] W. Chen, C. A. Choquette-Choo, P. Kairouz, and A. T. Suresh, "The fundamental price of secure aggregation in differentially private federated learning," in *International Conference on Machine Learning (ICML)*, 2022.

[17] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1, 2021.

[18] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *International Conference on Learning Representations*, 2020.

[19] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–9.

[20] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. on Neural Networks and Learning Sys.*, 2021.

[21] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *CoRR*, vol. abs/2002.10619, 2020.

[22] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Conf. on Neural Inf. Processing Sys. (NeurIPS)*, 2020.

[23] ——, "An efficient framework for clustered federated learning," *IEEE Trans. Inf. Theory*, 2022.

[24] Y. Ruan and C. Joe-Wong, "Fedsoft: Soft clustered federated learning with proximal local updating," in *AAAI Conf. on Artificial Intelligence*, 2022.

[25] M. Nafea, E. Shin, and A. Yener, "Proportional fair clustered federated learning," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, 2022.

[26] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[27] A. Fallah, A. Mokhtari, and A. E. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized federated learning with moreau envelopes," in *Advances in Neural Inf. Processing Sys. (NeurIPS)*, 2020.

[29] K. Singhal, H. Sidahmed, Z. Garrett, S. Wu, J. Rush, and S. Prakash, "Federated reconstruction: Partially local federated learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[30] J. So, R. E. Ali, B. Guler, J. Jiao, and S. Avestimehr, "Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, 2023.

[31] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," *arXiv:1802.07876*, 2018.

[32] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Machine Learning and Sys. (MLSys)*, 2020.

[33] J. Shu, T. Yang, X. Liao, F. Chen, Y. Xiao, K. Yang, and X. Jia, "Clustered federated multitask learning on non-iid data with enhanced privacy," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3453–3467, 2023.

[34] J. Shao, Y. Sun, S. Li, and J. Zhang, "Dres-fl: Dropout-resilient secure federated learning for non-iid clients via secret data sharing," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[35] K. S. Kedlaya and C. Umans, "Fast polynomial factorization and modular composition," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1767–1802, 2011.

[36] A. R. Elkordy, J. Zhang, Y. H. Ezzeldin, K. Psounis, and S. Avestimehr, "How much privacy does federated learning with secure aggregation guarantee?" *Proc. Priv. Enhancing Technol.*, 2023.

[37] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," 2010.

[38] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[39] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *26th Annual Network and Distributed Sys. Security Symp. (NDSS)*, 2019.

[40] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," *IEEE Symp. on Security and Privacy*, 2016.

[41] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE Symp. on Security and Privacy*, 2019, pp. 739–753.

[42] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE Journal on Sel. Areas in Comm.*, vol. 39, no. 7, pp. 2168–2181, 2021.

[43] P. Feldman, "A practical scheme for non-interactive verifiable secret sharing," in *28th Annual Symp. on Foundations of Comp. Science*, 1987, pp. 427–438.

[44] S. Gao, "A new algorithm for decoding reed-solomon codes," *Communications, Information and Network Security*, pp. 55–68, 2003.

[45] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[46] J.-E. Ekberg, K. Kostiainen, and N. Asokan, "Trusted execution environments on mobile devices," in *ACM Conf. on Comp. and Comm. Security (CCS)*, 2013.

[47] J. So, B. Güler, and S. Avestimehr, "A scalable approach

for privacy-preserving collaborative machine learning," in *Advances in Neural Inf. Proc. Sys. (NeurIPS)*, 2020.

[48] M. C. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.

[49] Q. Yu, S. Li, N. Raviv, S. M. M. Kalan, M. Soltanolkotabi, and S. A. Avestimehr, "Lagrange coded computing: Optimal design for resiliency, security, and privacy," in *Int. Conf. on Art. Int. and Stat.*, 2019.

Hasin Us Sami (Student Member, IEEE) received his B.Sc. degree in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2019. He is currently a Ph.D. student at the Department of Electrical and Computer Engineering, University of California, Riverside. His research interests include federated and distributed machine learning, information theory, secure and private computing, and wireless networks.

Basak Guler (Member, IEEE) received the B.Sc. degree in Electrical and Electronics Engineering from the Middle East Technical University (METU), Ankara, Turkey, and the Ph.D. degree from the Wireless Communications and Networking Laboratory at the Pennsylvania State University in 2017. From 2018 to 2020, she was a Postdoctoral Scholar at the University of Southern California. She is currently an Assistant Professor at the Department of Electrical and Computer Engineering, University of California, Riverside. Dr. Guler has received an NSF CAREER Award in 2022. Her research interests include information theory, distributed computing, machine learning, and wireless networks.