Rate-Distortion Optimization for Adaptive Gradient Quantization in Federated Learning

Guojun Chen*†, Lu Yu‡, Wenqiang Luo†, Yinfei Xu†, and Tiecheng Song*†
*National Mobile Communication Research Laboratory, Southeast University, Nanjing 210096, China
†School of Information Science and Engineering, Southeast University, Nanjing 210096, China
‡China Mobile Research Institute, Beijing 100053, China
Email: guojunchen@seu.edu.cn, yulu@chinamobile.com, luowenqiang_98@163.com, {yinfeixu, songtc}@seu.edu.cn

***Abstract*—Federated learning (FL) is an emerging machine learning setting designed to preserve privacy. However, constantly updating model parameters on uplink channels results in huge communication overload, which is a major challenge for FL. In this paper, we consider an adaptive gradient quantization approach based on rate-distortion optimization in FL, which consists of a non-stationary random walk model on the true global optimal model parameters. Unlike traditional quantization methods, our goal is to minimize the total communication costs when the global server reconstructs model parameters under distortion constraints. Furthermore, when considering the iterative process, we utilize the Kalman filter to reduce computational complexity. And in each iteration, a generalized water-filling algorithm is used to calculate the optimal quantization levels for each local client. Numerical results show that the proposed method outperforms conventional quantization methods in terms of reducing communication costs.**

***Index Terms*—Federated Learning, Communication Efficiency, Adaptive Quantization, Rate-Distortion.**

I. Introduction

Distributed learning has been widely used in wireless sensor networks and the Internet of Things to transmit data between local devices and a central server. As a famous learning model, federated learning (FL) allows multiple local clients to jointly train a machine learning model on their combined data, without revealing their data to a global server [1].

However, one of the major challenges of FL is that constantly updating model parameters can lead to a massive communication overload, which significantly slows down the convergence of FL [2]. In this paper, we consider a FL system with over-the-air computation over a multiple access channel (MAC) where local gradients are compressed before transmission to meet the bandwidth limitation [3]. One of the most popular methods is quantization, which involves lossy compression of gradients through quantizing entry to a finite-bit low precision value [4]. [5] firstly applied stochastic uniform quantizer on the gradient. [6] considered the correlation inside the gradient vector and raised the vector version of QSGD. Furthermore, [7] proposed an adaptive quantization strategy for each local client.

These works significantly reduce the communication burden of FL systems at the cost of degrading model performance or increasing computational complexity. Apart from introducing the drift of global model parameters in FL [8], which is named dynamic FL. [9] attempted to compute the rate region of the local encoders, which inspirited us to apply the lossy source coding theory. However, they only came to a weak conclusion and the model is hard to realize in real FL systems.

In order to reduce communication costs, it is preferable to send quantized parameters with minimized information entropy according to information theory. The Gaussian CEO (read either Chief Executive Officer or Central Estimation Officer) problem is a well-known model in lossy source coding theory for decades. A tight upper bound on the sum-rate distortion function was solved in [10] using the "Generalized Waterfilling" approach. The work in [11] extended the Gaussian CEO problem by tracking and then proposed a suboptimal water-filling allocation algorithm at last. Inspired by the Kalman filter utilized in causal source coding theory [12] and applied CEO system in distributed training problem, we propose the Rate-Distortion Optimization for Adaptive Gradient Quantization(RDO-AGQ) in FL system for communication efficiency. The main contributions of this paper are summarized as follows:

- We propose a novel RDO-AGQ approach for communication efficient FL. Different from [7] and [13], our approach calculates the adaptive quantization strategy online not only for each communication round but also for each local client.
- In order to reduce communication costs and preserve privacy, we introduce the rate-distortion theory into FL framework. Compared with conventional the FL framework, our rate-distortion optimization approach considers the drift of global model parameters and aims at minimizing the communication costs under distortion constraints. Different from distributed learning model [9], [14], [15], we consider the memory of local clients and a central server to improve communication efficiency.
- Numerical results validate the effectiveness of the proposed approach and show that this approach significantly reduces the upload communication costs when compared with conventional FL systems using the quantization approach, e.g. [5], [7].

The rest of this paper is organized as follows. Section II presents the preliminaries of the FL system. Section III

The first two authors make equal contributions to this work.

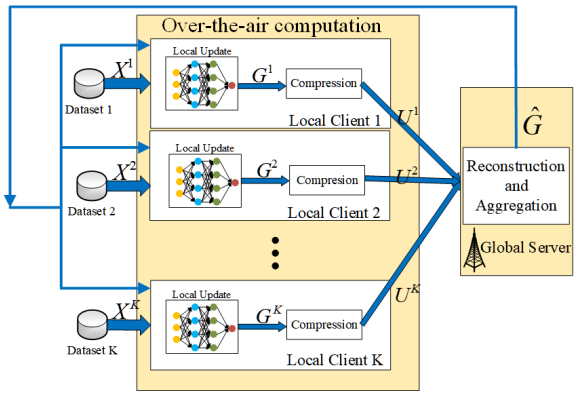

Fig. 1. Conventional federated learning framework

details the framework of RDO-AGQ FL and formulates the problem. The process of solving the rate optimization problem is proposed in Section IV. Section V provides the numerical results and Section VI concludes the paper.

Throughout the paper, we usually use capital letters (say, X) to indicate a random variable. $X^{[K]}$ denotes the random vector $(X^1, \ldots, X^K)$ in the clients' dimension. And $X_{[t]}$ denotes the random vector $(X_1, \ldots, X_t)$ in training rounds dimension.

II. Preliminary of Federated Learning

In this section, we present the necessary preliminaries about the FL framework. We first introduce the FL system with communication costs. Then, the stochastic uniform quantizer and communication costs are introduced and defined.

A. Federated Learning System with Communication Costs

As depicted in Fig.1, we consider a conventional FL framework in one training round, which consists of K local clients and one global server dispersed in space. The training data, distributed among local clients, is denoted as $\boldsymbol{X}^k := \{X^k\}_{i=1}^{n_k}$, which is the set of n_k labeled training samples available at the kth client for some $k \in \{1, 2, \ldots, K\}$. Each client has access to $\boldsymbol{X}^k$ and $\boldsymbol{G}$ to train the local model, where $\boldsymbol{G} := \{G_1, G_2, \ldots G_P\}$ is the gradient vector or learning model parameters averaged by the global server. In order to preserve the privacy of each local client, we consider $\boldsymbol{G}$ as the gradient vector in this paper, and local clients upload gradients instead of private model parameters using Stochastic Gradient Descent (SGD).

Then, the kth client trains a machine learning model to minimize the objective function $F^k(\cdot)$. which is defined as the empirical average over the corresponding training set i.e.,

$$F^k(\boldsymbol{G}; \boldsymbol{X}^k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(\boldsymbol{G}; X^k), \tag{1}$$

where $\ell(\cdot;\cdot)$ is the loss function. Then, local clients output model parameters $\boldsymbol{\omega}^{[K]} = \{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \ldots \boldsymbol{\omega}^K\}$ and transform them into gradient vectors by SGD, denoted as $\boldsymbol{G}^{[K]} = \{\boldsymbol{G}^1, \boldsymbol{G}^2, \ldots \boldsymbol{G}^K\}$.

After the local training phase, the updated local gradient vectors need to transmit to the global server through a separate finite-bit channel. Based on methods used in numerous researches [5]- [9]. We propose a novel gradient compassion approach for efficient communication in this paper, which will be described in subsequent sections.

By receiving all the compressed gradient vectors, defined as $\boldsymbol{U}^{[K]} := \{\boldsymbol{U}^1, \boldsymbol{U}^2, \cdots, \boldsymbol{U}^K\}$, from local clients, the global server reconstructs the local gradient vectors and estimates a global update $\hat{\boldsymbol{G}}$, which is the average of K unbiased estimators via

$$\hat{\boldsymbol{G}}^k = \mathbb{E}[\boldsymbol{G}^k | \boldsymbol{U}^{[K]}], \tag{2}$$

$$\hat{\boldsymbol{G}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{G}}^k, \tag{3}$$

where $\boldsymbol{G}^k$ and $\hat{\boldsymbol{G}}^k$ are the local true gradient vector and its unbiased estimate, which is the same as the definition in [6, Theorem 3]. At last, a new training iteration is started by broadcasting the globally updated gradient to all local clients.

B. Stochastic Uniform Quantizer and Communication Costs

In this section, we first introduce a commonly used [5], [7] quantitative strategy, which is also applied in this paper. Here, we repeat the stochastic uniform quantizer defined in [7, Section II].

Stochastic Uniform Quantizer: The stochastic uniform quantization operator $Q_s(\boldsymbol{\omega})$ is parameterized by the number of quantization levels $s \in \mathbb{N} = \{1, 2, \ldots\}$. For each dimension of a d-dimensional parameter vector $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_d]$,

$$Q_s(\omega_i) = ||\boldsymbol{\omega}||_2 \text{sign}(\omega_i) \eta_i(\boldsymbol{\omega}, s), \tag{4}$$

where $\eta_i(\omega, s)$ is a random variable given as,

$$\eta_i(\omega, s) = \begin{cases} \frac{l+1}{s} & \text{with probability } \frac{|\omega_i|}{||\boldsymbol{\omega}||_2} s - l \\ \frac{l}{s} & \text{otherwise} \end{cases} \tag{5}$$

Here, $l \in \{0, 1, 2, \ldots, s-1\}$ is an integer such that $\frac{|\omega_i|}{||\boldsymbol{\omega}||_2} \in \left[\frac{l}{s}, \frac{l+1}{s}\right)$. For $\boldsymbol{\omega} = \mathbf{0}$, we define $Q_s(\boldsymbol{\omega}) = \mathbf{0}$.

Communication Costs: The communication costs mentioned in this paper are defined as the number of bits communicated by local clients to the global server. At the first place, given $Q_s(\omega_i)$, we need 1 bit to represent $\text{sign}(\omega_i)$ and $\lfloor \log_2(s+1) \rfloor$ bits to represent $\eta_i(\boldsymbol{\omega}, s)$. The scalar $||\boldsymbol{\omega}||_2$ to be 32 bits. Thus, the number of bits communicated by the kth local client to the global serve per round, which we denoted by C_s^k, is given by

$$C_s^k = d\lfloor \log_2(s+1) \rfloor + d + 32. \tag{6}$$

Therefore, the proposed RDO-AGQ algorithm is aimed at minimizing the sum communication costs per iterative round, which is denoted as $C_\Sigma = \sum_{k=1}^{K} C_s^k$, under a distortion constraint.

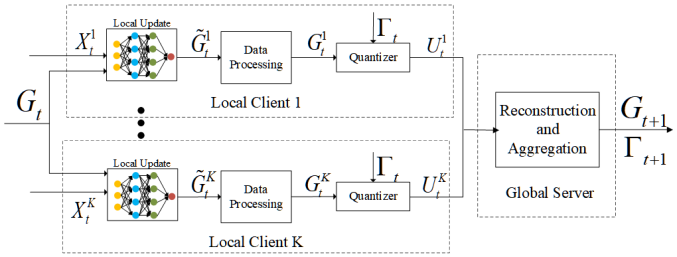

Fig. 2. Framework of RDO-AGQ FL

III. Framework of RDO-AGQ FL and Optimization Problem

This section presents the framework of RDO-AGQ FL system. And a rate-distortion optimization problem is proposed under some assumptions.

A. Framework of RDO-AGQ FL

In a conventional FL system, local clients only consider their own gradient vector of model parameters and upload it with a simple compressed method. The relevance between datasets and gradients is ignored. In this paper, we utilize the correlation of gradients in two aspects. First of all, referring to the distributed source coding system, we quantize the gradients in a distributed way. Secondly, because the FL system is a learning process, the gradients are related to that in previous rounds. We consider a dynamic FL system to compress gradients for improving communication efficiency. Then, we are ready to establish our RDO-AGQ FL framework.

As seen in Fig 2, the gradient vector computed by the kth local client is denoted as $\tilde{\boldsymbol{G}}_t^k$. For convenient computation, we apply the data processing method mentioned in [15] to transform the gradient vectors $\tilde{\boldsymbol{G}}_t^{[K]}$ into correlated Gaussian random variables $\boldsymbol{G}_t^{[K]}$.

Then, local clients aim at quantizing the processed gradient vector $\boldsymbol{G}_t^{[K]}$ with minimized summed communication costs C_Σ based on the distortion constraints between true and estimated global gradient vector. However, in the FL system, a local client is unable to have access to private information from other clients. The distributed source coding theory is hard to work on here. In our RDO-AGQ FL system, we require the global server broadcast a vector $\boldsymbol{\Gamma}_t$ which is the function of the whole variance of the unbiased estimate, defined in (27).

To make the communication more efficient and let the local clients compute the optimal $\boldsymbol{s}_t^k$ only with $\boldsymbol{\Gamma}_t$, we introduce the dynamic FL system. Assumed that the local clients and global server have memory of all previous information, the kth local client calculates the optimal $\boldsymbol{s}_t^k$ with $\boldsymbol{G}_{[t]}^k$, $\boldsymbol{U}_{[t-1]}^k$ and $\boldsymbol{\Gamma}_t$, while the global server estimates the processed gradient vectors with $\boldsymbol{U}_{[t]}^{[K]}$, where $\boldsymbol{U}_t^k$ is the codeword for compressed gradient vector. The detailed methods to compute the optimal quantization level $\boldsymbol{s}_t^k$ require the assumption of dynamic FL and the relevance of gradients in the time dimension, which would be discussed in subsequent content.

After determining each optimal quantized level by local clients, the quantized gradient vectors are uploaded to the global server with the rate R_t^k, which is defined as $R_t^k := \log|\boldsymbol{s}_t^k|$. Therefore, it is equal to finding the optimal communication rate and quantized level. With the method of variable length code, it is valid for the communication rate R_t^k being non-integer.

The global server constructs and arranges the received processed gradient vectors using (2) and (3). Then, it broadcasts the gradient vector and $\boldsymbol{\Gamma}_t$ to all the local clients to finish a training round. The pseudo-code of the training process is described in the following algorithm.

Algorithm 1 Algorithm of RDO-AGQ FL system

Input: Total number of clients K; Total number of communication rounds T; Local dataset $(\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^K)$; The set of distortion constrain $(D_1, D_2, \ldots, D_T)$;

Initialization: Global model gradient vector $\boldsymbol{G}_0$;
for $t = 1$ to T **do**
  for $k = 1$ to K **do**
    Train the model for minimizing $F_t^k(\boldsymbol{G}; \boldsymbol{X}^k)$ in (1)
    Calculate current gradient of model parameters $\tilde{\boldsymbol{G}}_t^k$
    Transform $\tilde{\boldsymbol{G}}_t^k$ to $\boldsymbol{G}_t^k$ using data processing method.
    Compute the optimal quantized level with $\boldsymbol{G}_{[t]}^k$, $\boldsymbol{U}_{[t-1]}^k$ and $\boldsymbol{\Gamma}_t$ using RDO-AGQ algorithm.
    Quantize the processed gradient vector $\boldsymbol{G}_t^k$ and transmit to the global server.
  end for
  Reconstructs $\boldsymbol{G}_t^k$ by unbiased estimator (2).
  Post-process $\boldsymbol{G}_t$ and arrange the global updated gradient vector $\tilde{\boldsymbol{G}}_t$ by (3).
  Broadcast the updated gradient vector and $\boldsymbol{\Gamma}_t$ to all local clients.
end for

Next, we are willing to establish a method for computing the optimal quantized level.

B. Optimization Problem

The goal of RDO-AGQ is to find the optimal communication rates R_t^k to minimize communication costs C_Σ for all uplink channels with distortion constraints. To faithfully represent the FL setup, we design our RDO-AGQ strategy in light of the following requirements and assumptions: Firstly, we impose an assumption on the drift of the global updated gradient vector $\boldsymbol{G}_t$, namely that it follows a random walk model as in [8].

Assumption 1: We assume that the global updated gradient vector $\boldsymbol{G}_t$ follows a random walk:

$$\boldsymbol{G}_t = \boldsymbol{C} * \boldsymbol{G}_{t-1} + \boldsymbol{W}_{t-1} \tag{7}$$

where $\boldsymbol{W}_t$ denotes some zero mean random vectors independent of $\boldsymbol{G}_{t'}$ for any $t' < t$ and with bounded variance, i.e., $\mathbb{E}[||\boldsymbol{W}_t||^2] = \sigma^2_{\boldsymbol{W}_t}$. Under the Gaussian cases in this paper, let the independent random vector $\boldsymbol{W}_t$ is distributed as $\mathcal{N}(0, \sigma^2_{\boldsymbol{W}_t})$.

Then, the estimate of the optimum parameters by the kth local client would be noisily referred to [14]. And due to all the variables $\boldsymbol{G}_t$ and $\boldsymbol{G}_t^k$ are zero-mean Gaussian random vectors, we can make the following assumption

Assumption 2: We assume the local updated gradient vector $\boldsymbol{G}_t^k$ is modeled similar as [14, Eq.(1)] via

$$\boldsymbol{G}_t^k = \boldsymbol{G}_t + \boldsymbol{N}_t^k, \tag{8}$$

where $\{\boldsymbol{N}_t^1, \ldots, \boldsymbol{N}_t^K\}_{t=1}^T$ are Gaussian random variables independent of $\boldsymbol{G}_t$ with independent components and each component of $\boldsymbol{N}_t^k$ is distributed as $\mathcal{N}(0, \sigma^2_{\boldsymbol{N}_k})$.

At last, we consider the distortion between the true gradient $\boldsymbol{G}_t$ and its unbiased estimate $\hat{\boldsymbol{G}}$, computed at the global server by (3), is under a quadratic distortion measure. This assumption is widely used such as [14], [6, Section III], et.al.

Assumption 3: The main measure of distortion is following the mean squared error, via

$$D(\boldsymbol{G}_t, \hat{\boldsymbol{G}}_t) = \frac{1}{K}\mathbb{E}\left\{\left(\boldsymbol{G}_t^k - \hat{\boldsymbol{G}}_t^k\right)^2\right\}. \tag{9}$$

At last, our RDO-AGQ approach aims to calculate the minimum communication costs C_Σ under given distortion contractions D_t per iterative round. According to the definition of C_Σ in (6) and the fact of $R_t^k := \log|\boldsymbol{s}_t^k|$, it is equal to calculate the minimized sum rates of $R_t^{[K]}$ and C_Σ. Thus, we propose the problem formulation via:

$$(P1) \quad R_t(D_t) = \min_{\boldsymbol{U}_t^K} \sum_{k=1}^K R_t^k \tag{10}$$

$$s.t. \quad \mathbb{E}[||\boldsymbol{G}_t^k - \hat{\boldsymbol{G}}_t^k||^2] \le D_t, \tag{11}$$

where, the unbiased estimate $\hat{\boldsymbol{G}}_t$ follows (3). In light of the above assumptions and problem formulation, we propose a solution to problem P1 in the next section.

IV. Solving the Rate Optimization Problem P1

In this section, we propose an algorithm to solve the above problem P1 and calculate the optimal communication rate for each local client.

A. Conversion Optimization Problem P1 with Kalman Filter

In this section, we converse the optimal problem P1 into a calculable form P2 by Kalman Filter. At the first place, let us make some notations here. Repeat (2), the unbiased estimate of processed local gradient vector $\boldsymbol{G}_t^k$ is

$$\hat{\boldsymbol{G}}_t^k = \mathbb{E}[\boldsymbol{G}_t|\boldsymbol{U}_{[t]}^k]. \tag{12}$$

And let the prediction of $\boldsymbol{G}_t^k$, using $\boldsymbol{G}_{[t]}^k$ in past $t-1$ iterative rounds, be $\hat{\boldsymbol{G}}_{t|t-1}^k$, via:

$$\hat{\boldsymbol{G}}_{t|t-1}^k = \mathbb{E}[\boldsymbol{G}_t|\boldsymbol{U}_{[t-1]}^k]. \tag{13}$$

Moreover, let the error variance between the true global gradient vector $\boldsymbol{G}_t$ and the kth unbiased estimate of local gradient vector $\hat{\boldsymbol{G}}_t^k$ be $\boldsymbol{P}_t^k$, via:

$$\boldsymbol{P}_t^k = \mathbb{E}[(\boldsymbol{G}_t - \hat{\boldsymbol{G}}_t^k)^2]. \tag{14}$$

And let the error variance between $\boldsymbol{G}_t^k$ and the predicted gradient vector $\hat{\boldsymbol{G}}_{t|t-1}^k$ be $\boldsymbol{P}_{t|t-1}^k$, via:

$$\boldsymbol{P}_{t|t-1}^k = \mathbb{E}[(\boldsymbol{G}_t - \hat{\boldsymbol{G}}_{t|t-1}^k)^2]. \tag{15}$$

Furthermore, according to Assumption 2, we define the estimate $\boldsymbol{G}_t$ by the kth local client as $\overline{\boldsymbol{G}}_t^k := \mathbb{E}[\boldsymbol{G}_t|\boldsymbol{G}_{[t]}^k]$ for simple derivation. Similar as (13), $\overline{\boldsymbol{G}}_{t|t-1}^k := \mathbb{E}[\boldsymbol{G}_t|\boldsymbol{G}_{[t-1]}^k]$ is defined as the predicted estimate on $\boldsymbol{G}_t$ by previous information on the kth client. Symmetrically, the error variance between $\boldsymbol{G}_t$ with $\overline{\boldsymbol{G}}_t^k$ and $\overline{\boldsymbol{G}}_{t|t-1}^k$ are respectively defined as

$$\boldsymbol{Q}_t^k = \mathbb{E}[(\boldsymbol{G}_t - \overline{\boldsymbol{G}}_t^k)^2] \tag{16}$$

$$\boldsymbol{Q}_{t|t-1}^k = \mathbb{E}[(\boldsymbol{G}_t - \overline{\boldsymbol{G}}_{t|t-1}^k)^2] \tag{17}$$

Using the above definitions and defining the variance $\boldsymbol{G}_t$ as $\sigma^2_{\boldsymbol{G}_t}$, we now present the conversion of problem P1 which aims to calculate the minimized communication rate.

Theorem 1: The rate-distortion optimization in this system model is equal to the following optimal problem.

$$(P2) \quad R_t(D_t) = \inf_{\boldsymbol{P}_t^{[K]}} \frac{1}{2}\log\frac{\sum_{k=1}^K \frac{1}{\boldsymbol{P}_t^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}}}{\sum_{k=1}^K \frac{1}{\boldsymbol{P}_{t|t-1}^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}}} + \sum_{k=1}^K \frac{1}{2}\log\frac{(1-\frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t-1}^k})}{(1-\frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t}^k})}. \tag{18}$$

$$s.t. \quad \boldsymbol{P}_{t|t}^k > \boldsymbol{Q}_t^k \tag{19}$$

$$\boldsymbol{P}_{t|t}^k \ge (\frac{1}{\sigma^2_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k})^{-1} \tag{20}$$

$$\sum_{k=1}^K (\frac{1}{\boldsymbol{P}_{t|t-1}^k} - \frac{1}{\sigma^2_{\boldsymbol{G}_t}}) + \frac{1}{\sigma^2_{\boldsymbol{G}_t}} \ge \frac{1}{D_t} \tag{21}$$

with $\boldsymbol{P}_{t|t-1}^k$ and $\boldsymbol{P}_t^k$ satisfying

$$\boldsymbol{P}_{t|t-1}^k = a_{t-1}^2 \boldsymbol{P}_{t-1} + \sigma^2_{W_t}, \tag{22}$$

$$\boldsymbol{P}_t^k = \Big(\frac{1}{\sigma^2_{\boldsymbol{N}_t^k} + \frac{\sigma^2_{\boldsymbol{V}_t^k}}{(h_t^k)^2}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k}\Big)^{-1}. \tag{23}$$

Where, $\boldsymbol{V}_t^k$ and h_t^k are the auxiliary variables.

Proof: Here, we proposed the proof of this theorem briefly and the detailed proof is proposed in Appendix A. Firstly, the communication rate R_t^k on each local client would be expended using mutual information via:

$$R_t^k \ge I(\boldsymbol{G}_{[t]}^{[K]}; \boldsymbol{U}_t^{[K]}|\boldsymbol{U}_{[t-1]}^{[K]}) \tag{24}$$

Afterward, the Kalman filter would be utilized for estimating the global updated gradient vector $\boldsymbol{G}_t$ using observing data before quantizers. Jointing the Kalman filter result with [11, Theorem 4], this theorem is easy to prove by a simple operation. ∎

B. Solving the Optimization Problem P2

In this section, we propose an iterative water-filling algorithm to solve the above optimal problem P2 in Theorem 1 at each iterative round t. Because the error covariance $\boldsymbol{P}_{t|t-1}^k$ is calculated by the time slot $t-1$, this time we only need to find the optimal $\boldsymbol{P}_t^k$. To calculate the minimum communication costs is equivalent to finding the optimal error covariance of estimate $\boldsymbol{P}_t^k$.

Algorithm 2 Iterative Waterfilling Algorithm

1: **Initialization**: $\boldsymbol{P}_t^k$ for $k = 1, 2, ..., K$, ν_t and $\Delta R = \infty$.
2: **while** $\Delta R > \epsilon$ **do**
3: **for** k = 1 to K **do**
4: Update $\tilde{\boldsymbol{\Gamma}}_t^k = \boldsymbol{\Gamma}_t - \frac{1}{\boldsymbol{P}_t^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}}$.
5: Update objective parameter $\boldsymbol{P}_t^k$ with (25)
6: Update communication rate threshold ν_t with (26)
7: **end for**
8: Update $R_t(D_t)$ with (18)
9: **end while**

where

$$\frac{1}{\boldsymbol{P}_{t|t}^k} = \frac{1}{2}\left[(\frac{1}{\boldsymbol{Q}_t^k} - \tilde{\boldsymbol{\Gamma}}_t^k) + \sqrt{(\frac{1}{\boldsymbol{Q}_t^k} + \tilde{\boldsymbol{\Gamma}}_t^k)^2 - \frac{4(\frac{1}{\boldsymbol{Q}_t^k} + \tilde{\boldsymbol{\Gamma}}_t^k)}{\nu_t}}\right] \quad (25)$$

$$\nu_t = \arg\left\{\sum_{k=1}^{K}(\frac{1}{\boldsymbol{P}_t^k} - \frac{1}{\sigma^2_{\boldsymbol{G}_t}}) + \frac{1}{\sigma^2_{\boldsymbol{G}_t}} = \frac{1}{D_t}\right\}. \quad (26)$$

$$\boldsymbol{\Gamma}_t = \sum_k (\frac{1}{\boldsymbol{P}_t^j}) \quad (27)$$

(28)

The iterative waterfilling algorithm is presented in Algorithm 2. From this algorithm, $R_t(D_t)$ is the direct output, and the communication rates $R_t^{[K]}$ for local clients are easy to be operated by threshold ν_t. This algorithm is an expanded form of the iterative water-filling algorithm. And the theoretical analysis is proposed in Appendix B.

V. Numerical Results

In this section, we numerically evaluate our RDO-AGQ approach. We show the performance of reducing communication rate using the proposed RDO-AGQ by simulations.

A. Evaluations Settings

We consider a conventional FL system with one server and K clients and distribute datasets to each of the clients. The settings are listed in the following.

(1) Dataset: We use the MNIST dataset and CIFAR-10 dataset in the experiments.

(2) Data Distributions: We consider both IID and non-IID data distributions in this paper. For the IID dataset partition, the data samples are uniformly randomly assigned to clients. For the non-IID dataset partition, the data samples are sorted by their labels and divided into $2n$ groups, and each client receives two groups.

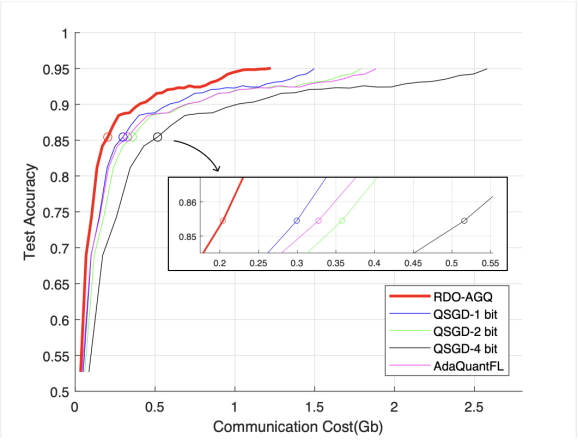

Fig. 3. Test accuracy versus communication costs. IID partition on MNIST.

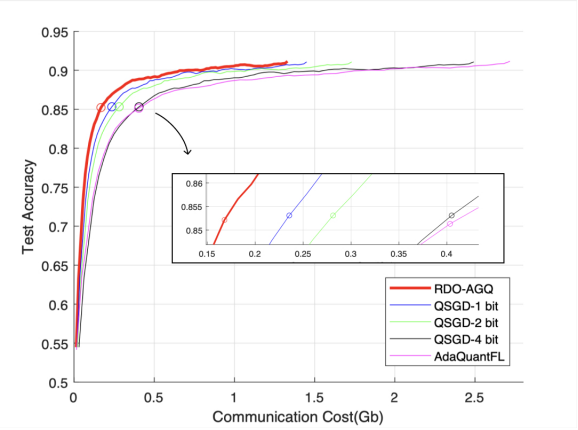

Fig. 4. Test accuracy versus communication costs. IID partition on CIFAR-10.

(3) Basic Configuration: We use the following configuration in our experiments.

- Total amount of clients: K = 1000.
- Learning rate: $\gamma = 0.001$
- Quantizer: Stochastic uniform quantizer [5], [7].
- Local updating uses CNN architecture for MNIST and Resnet-56 architecture for CIFAR-10.
- Local clients train the model for 10 epochs.
- Compared Algorithms:QSGD [5], AdaQuantFL [7].

B. Performance of RDO-AGQ

In this section, we consider the performance of the proposed RDO-AGQ approach with MNIST dataset and CIFAR-10 dataset. In general, RDO-AGQ only needs fewer total communication costs to achieve specific test accuracy. And with the same communication costs, the FL system is able to achieve higher test accuracy using RDO-AGQ.

In general, as seen in Fig. 1-4, RDO-AGQ only needs fewer total communication costs to achieve specific test accuracy. And with the same communication costs, the FL system is able to achieve higher test accuracy using RDO-AGQ. In detail, described in Fig. 1, when the dataset is i.i.d

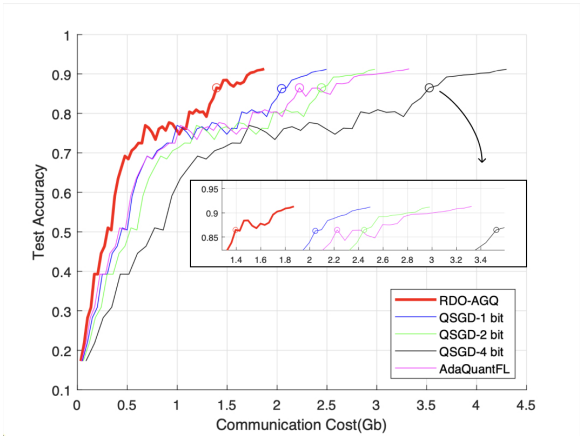

Fig. 5. Test accuracy versus communication costs. Non-IID partition on MNIST.

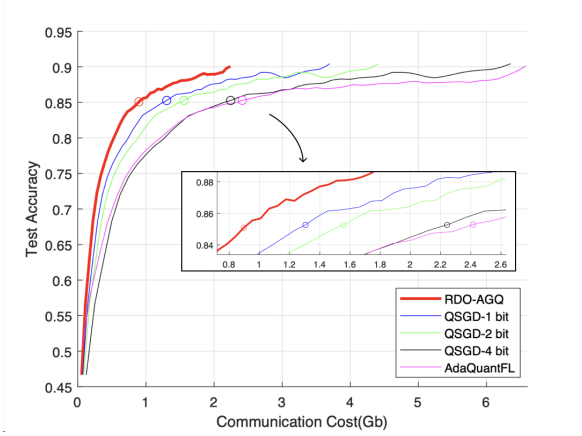

Fig. 6. Test accuracy versus communication costs. Non-IID partition on CIFAR-10.

distributed MNIST dataset, RDO-AQG only demands 204Mb and 1224Mb communication costs to reach the test accuracy of 85% and 95%, respectively. However, to achieve the same test accuracy, the next best method is the 1-bit QSGD method which costs approximately 299Mb and 1497Mb, respectively. Therefore, our proposed RDO-AQG approach can save about 31.8% and 18.2% communication costs for the same test accuracy, respectively. Fig. 2 describes the relevance of test accuracy and communication costs for i.i.d CIFAR-10 dataset. It shows that RDO-AQG only demands 168Mb and 1330Mb communication costs to reach the test accuracy of 85% and 91%, respectively. While the next best method, achieving the same test accuracy, costs approximately 235Mb and 1449Mb, respectively. Thus, the RDO-AQG approach can save about 28.5% and 8.2% communication costs for the same test accuracy, respectively. Similarly, Fig. 3 shows that RDO-AQG approach can save about 31.8% and 25.0% communication costs to reach the test accuracy of 85% and 91% for non-i.i.d. MNIST dataset. And Fig. 4 shows that RDO-AQG approach can save about 31.2% and 39.4% communication costs to reach the test accuracy of 85% and 90% for non-i.i.d. MNIST.

In conclusion, compared to AdaQuantFL and QSGD with different fixed quantization levels, the performance of the proposed RDO-AGQ approach is the best for MNIST and CIFAR-10 datasets in the FL system. RDO-AGQ approach can significantly reduce communication costs and shows robustness during simulation. These lower communication costs are manifested through using less precious communication sources, which are scarce in upload channels.

VI. Conclusion

This paper introduces an RDO-AGQ approach for communication-efficient FL. We apply lossy source coding theory to solve the optimal quantization strategy. Here the Kalman filter and generalized water-filling algorithm are used to calculate the optimal quantization levels for each local client and each label dimension of gradients. Numerical results show that the RDO-AGQ outperforms AdaQuanFL and QSGD.

Appendix A
Proof of Theorem 1

Firstly, to prove Theorem 1, a necessary lemma is presented here for subsequent proof.

Lemma 1: [11, Appendix D] We let Gaussian random variable $X \sim \mathcal{N}(0, \sigma_X^2)$ with

$$Y_k = X + W_k, \quad k = 1, \ldots, K, \tag{29}$$

where $W_k \sim \mathcal{N}(0, \sigma_{W_k}^2)$ with $W_k \perp W_j$ for $j \neq k$. Then the MMSE estimate and the normalized estimate error of X given $Y_{[K]}$ are given by

$$\mathbb{E}[X|Y_{[K]}] = \sum_{k=1}^{K} \frac{\sigma^2_{X|Y_{[K]}}}{\sigma^2_{W_k}} Y_k, \tag{30}$$

$$\frac{1}{\sigma^2_{X|Y_{[K]}}} = \frac{1}{\sigma^2_X} + \sum_{k=1}^{K} \frac{1}{\sigma^2_{W_k}}. \tag{31}$$

Proof: The proof of this lemma is detailed in [11, Appendix D]. ∎

Then, using mutual information, problem P1 in (10) comes to the rate-distortion function

$$R_t(D_t) = \inf I(\boldsymbol{G}_{[t]}^{[K]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) \tag{32}$$

$$\overset{(a)}{=} \inf I(\overline{\boldsymbol{G}}_{[t]}^{[K]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) \tag{33}$$

$$\overset{(b)}{=} \inf I((\overline{\boldsymbol{G}}_{[t]}^{[K]}, \boldsymbol{G}_{[t]}); \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) \tag{34}$$

$$\overset{(c)}{=} \inf I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) \tag{35}$$

$$+ \sum_{k=1}^{K} I(\overline{\boldsymbol{G}}_{[t]}^{k}; \boldsymbol{U}_t^{k} | \boldsymbol{U}_{[t-1]}^{k}, \boldsymbol{G}_{[t]}) \tag{36}$$

(a) holds because of $\overline{\boldsymbol{G}}_t^k = \mathbb{E}[\boldsymbol{G}_t | \boldsymbol{G}_{[t]}^k]$ is the function of $\boldsymbol{G}_{[t]}^k$. (b) holds because of the chain rule of mutual information using with $I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}, \overline{\boldsymbol{G}}_{[t]}^{[K]}) = 0$. Equation (c) is from the fact that $P_{\boldsymbol{U}_t^{[K]} | \boldsymbol{G}_{[t]}^{[K]}, \boldsymbol{U}_{[t-1]}^{[K]}} = \prod_{k=1}^{K} P_{\boldsymbol{U}_t^k | \boldsymbol{G}_{[t]}^k, \boldsymbol{U}_{[t-1]}^k}$.

Afterwards, Kalman filter is used to estimate the true global gradient vector $\boldsymbol{G}_t$ in iterative round t by the each kth local

client with $\boldsymbol{G}_{[t]}^k$. Next theorem is the result of calculating the estimate local clients before coding.

Theorem 2: The estimate of true global updated model gradient $\boldsymbol{G}_t$ by the k local client is

$$\overline{\boldsymbol{G}}_t^k = \frac{a * \sigma^2_{\boldsymbol{N}_t^k}}{\boldsymbol{Q}_{t|t-1}^k + \sigma^2_{\boldsymbol{N}_t^k}} \overline{\boldsymbol{G}}_{t-1}^k + \frac{\boldsymbol{Q}_{t|t-1}^k}{\boldsymbol{Q}_{t|t-1}^k + \sigma^2_{\boldsymbol{N}_t^k}} \tilde{\boldsymbol{G}}_t^k \tag{37}$$

$$= a'^k_{t-1} \overline{\boldsymbol{G}}_{t-1}^k + \boldsymbol{W}'^k_{t-1} \tag{38}$$

which follows $\boldsymbol{Q}_t^k = \big(\frac{1}{\boldsymbol{Q}_{t|t-1}^k} + \frac{1}{\sigma^2_{\boldsymbol{N}_t^k}}\big)^{-1}$ and $\boldsymbol{Q}_{t|t-1}^k = a^2 * \boldsymbol{Q}_{t-1}^k + \sigma^2_{\boldsymbol{N}_t^k}$.

Proof: This proof only need to use the standard Kalman Filer method. Therefore

$$\overline{\boldsymbol{G}}_{t|t-1}^k = a * \overline{\boldsymbol{G}}_{t-1} \tag{39}$$

$$\boldsymbol{Q}_{t|t-1}^k = a^2 \boldsymbol{Q}_{t-1}^k + \sigma^2_{\boldsymbol{W}_{t-1}} \tag{40}$$

$$\overline{\boldsymbol{G}}_t^k = \overline{\boldsymbol{G}}_{t|t-1}^k + K^k_{\overline{\boldsymbol{G}}_t} \mu^k_{\overline{\boldsymbol{G}}_t} \tag{41}$$

$$\boldsymbol{Q}_t^k = (1 - K^k_{\overline{\boldsymbol{G}}_t}) \boldsymbol{Q}_{t|t-1}^k \tag{42}$$

With the innovation $p\boldsymbol{\mu}^k_{\overline{\boldsymbol{G}}_t} = \boldsymbol{G}_t^k - \overline{\boldsymbol{G}}_{t|t-1}^k = \overline{\boldsymbol{G}}_{t-1}^k - \overline{\boldsymbol{G}}_{t|t-1}^k + \boldsymbol{N}_t^k$ and gain of Kalman Filter $K^k_{\overline{\boldsymbol{G}}_t} = \boldsymbol{Q}_{t|t-1}^k (\boldsymbol{Q}_{t|t-1}^k + \sigma^2_{\boldsymbol{N}_t^k})^{-1}$ Then, this theorem is proved with some simple operation. ∎

Next, we focus on calculating (35). The difference of this function with the standard Kalman Filter form is that it has K layers of $\boldsymbol{U}_t^{[K]}$, which is seen as the observers. According to the problem formulation in Section III-B, let

$$\boldsymbol{U}_t^k = h_t^k \boldsymbol{G}_t^k + \boldsymbol{\Phi}_t^k \boldsymbol{U}_{[t-1]}^k + \boldsymbol{\Psi}_t^k \boldsymbol{G}_{[t-1]}^k + \boldsymbol{V}_t^k \tag{43}$$

$$= h_t^k \boldsymbol{G}_t + \boldsymbol{\Phi}_t^k U_{[t-1]}^k + \boldsymbol{\Psi}_t^k \boldsymbol{G}_{[t-1]}^k + \boldsymbol{Z}_t^k, \tag{44}$$

where h_t^k is a auxiliary factor, $\boldsymbol{V}_t^k$ is a Gaussian random variable independent of $(\boldsymbol{G}_t^k,\ \boldsymbol{U}_{[t-1]}^k,\ \boldsymbol{V}_t^j)$, $j \neq k$ and $\boldsymbol{Z}_t^k = \boldsymbol{V}_t^k + h_t^k \boldsymbol{N}_t^k + \boldsymbol{\Psi}_t^k \boldsymbol{N}_{[t-1]}^k$ is also Gaussian random variable independent of $(\boldsymbol{G}_t^k,\ \boldsymbol{U}_{[t-1]}^k,\ \boldsymbol{Z}_t^j)$, $j \neq k$. It is obvious to find the value of $\boldsymbol{\Phi}_t^k$ and $\boldsymbol{\Psi}_t^k$ has no effect of our objective $I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]})$. So, let $\boldsymbol{\phi}_t^k = \boldsymbol{\Psi}_t^k = 0$ to get

$$\boldsymbol{U}_t^k = h_t^k \boldsymbol{G}_t + \boldsymbol{Z}_t^k. \tag{45}$$

where, $\boldsymbol{Z}_t^k = \boldsymbol{V}_t^k + h_t^k \boldsymbol{N}_t^k$ now. We estimate the desired global updated model gradient $\boldsymbol{G}_t$ by each local encoded codeword using Kalman Filter. The filter form of the estimate is

$$\hat{\boldsymbol{G}}_{t|t-1}^k = a * \hat{\boldsymbol{G}}_{t-1} \tag{46}$$

$$\boldsymbol{P}_{t|t-1}^k = a^2 * \boldsymbol{P}_{t-1} + \sigma^2_{\boldsymbol{W}_t} \tag{47}$$

$$\hat{\boldsymbol{G}}_t^k = \hat{\boldsymbol{G}}_{t|t-1}^k + K^k_{\hat{\boldsymbol{G}}_t} \boldsymbol{\mu}_t^k \tag{48}$$

$$\boldsymbol{P}_t^k = (1 - h_t^k K^k_{\hat{\boldsymbol{G}}_t}) \boldsymbol{P}_{t|t-1}^k = \big(\frac{(h_t^k)^2}{\sigma_{\boldsymbol{Z}_t^{k^2}}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k}\big)^{-1} \tag{49}$$

With the gain of Kalman Filter as $K^k_{\hat{\boldsymbol{G}}_t} = h_t^k \boldsymbol{P}_{t|t-1}^k ((h_t^k)^2 \boldsymbol{P}_{t|t-1}^k + \sigma^2_{\boldsymbol{Z}_t^k})^{-1}$ The iterative form of (35) is shown in the following theorem.

Theorem 3: The iterative form of $I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]})$ is

$$I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) = \frac{1}{2} \log \frac{\sum_{k=1}^K \frac{1}{\boldsymbol{P}_{t|t}^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}}}{\sum_{k=1}^K \frac{1}{\boldsymbol{P}_{t|t-1}^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}}} \tag{50}$$

Proof: Firstly, we extend (35) as

$$I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]}) = \frac{1}{2} \log \frac{Cov(\boldsymbol{G}_t | \boldsymbol{U}_{[t-1]}^{[K]})}{Cov(\boldsymbol{G}_t | \boldsymbol{U}_{[t]}^{[K]})} \tag{51}$$

using the characterization of mutual information for Gaussian random variables, where the function $Cov(\cdot)$ is covariance matrix. According to Lemma 1, the covariance in (51) is

$$Cov^{-1}(\boldsymbol{G}_t | \boldsymbol{U}_{[t-1]}^{[K]}) = \sum_{k=1}^K \frac{1}{\boldsymbol{P}_{t|t-1}^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}} \tag{52}$$

$$Cov^{-1}(\boldsymbol{G}_t | \boldsymbol{U}_{[t]}^{[K]}) = \sum_{k=1}^K \frac{1}{\boldsymbol{P}_{t|t}^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t}} \tag{53}$$

This theorem is proved putting (52) and (53) into (51). ∎

At last, we are interested in calculating (36) using the parameters of $\boldsymbol{P}_t^k$ and $\boldsymbol{P}_{t|t-1}^k$. Because of the mutual information in this term is the summary of all K local clients, we can expend the mutual information of just the kth local client for simplify. Using the similar method above, via

$$\boldsymbol{U}_t^k = h_t^k \boldsymbol{G}_t^k + \boldsymbol{\Phi}_t^k \boldsymbol{U}_{[t-1]}^k + \boldsymbol{\Psi}_t^k \boldsymbol{X}_{[t-1]}^k + \boldsymbol{V}_t^k \tag{54}$$

$$= f_t^k \overline{\boldsymbol{G}}_t^k + \boldsymbol{\Phi}_t^k \boldsymbol{U}_{[t-1]}^k + \boldsymbol{\Psi}_t^k \boldsymbol{G}_{[t]} + \boldsymbol{R}_t^k, \tag{55}$$

where the $1 \times t$ vector $\boldsymbol{\Phi}_t^k = [\boldsymbol{\Psi}_t^k, h_t^k] - f_t^k \cdot \Sigma_{\boldsymbol{G}_t, \boldsymbol{G}_{[t]}^k} \cdot \Sigma^{-1}_{\boldsymbol{G}_{[t]}^k, \boldsymbol{G}_{[t]}^k}$, $\boldsymbol{R}_t^k = \boldsymbol{V}_t^k + [\boldsymbol{\Psi}_t^k, h_t^k] \cdot \boldsymbol{N}_{[t]}^k$ and f_t^k is the auxiliary variable. While the value of $\boldsymbol{\Phi}_t^k$ and $\boldsymbol{\Gamma}_t^k$ makes no effect on $I(\overline{\boldsymbol{G}}_t^k; \boldsymbol{U}_t^k | \boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t]})$. Therefore, we let $\boldsymbol{\Phi}_t^k = \boldsymbol{\Gamma}_t^k = 0$ to get $\boldsymbol{U}_t^k = f_t^k \overline{\boldsymbol{G}}_t^k + \boldsymbol{R}_t^k$ and $\boldsymbol{G}_t = \overline{\boldsymbol{G}}_t^k + \boldsymbol{E}_t^k$, where $\sigma^2_{\boldsymbol{E}_t^k} = \boldsymbol{Q}_t^k$. Moreover, $\boldsymbol{R}_t^k = \boldsymbol{Z}_t^k = \boldsymbol{V}_t^k + h_t^k \cdot \boldsymbol{N}_t^k$ now. The iterative form of (36) is shown in the following theorem.

Theorem 4: The iterative form of $I(\boldsymbol{G}_{[t]}; \boldsymbol{U}_t^{[K]} | \boldsymbol{U}_{[t-1]}^{[K]})$ is

$$I(\overline{\boldsymbol{G}}_{[t]}^k; \boldsymbol{U}_t^k | \boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t]}) = \frac{1}{2} \log \frac{(1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t-1}^k})}{(1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t}^k})} \tag{56}$$

Proof: Firstly, we extend (36) as

$$I(\overline{\boldsymbol{G}}_{[t]}^k; \boldsymbol{U}_t^k | \boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t]}) = \frac{1}{2} \log \frac{Cov(\overline{\boldsymbol{G}}_t^k | \boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t]})}{Cov(\overline{\boldsymbol{G}}_t^k | \boldsymbol{U}_{[t]}^k, \boldsymbol{G}_{[t]})} \tag{57}$$

One of the nontrivial point here is that for the standard Kalman Filter method, it can only deal with the variables as $\mathbb{E}[\overline{\boldsymbol{G}}_t^k | \boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t-1]}]$ and $\mathbb{E}[\overline{\boldsymbol{G}}_t^k | \boldsymbol{U}_{[t]}^k, \boldsymbol{G}_{[t]}]$ which is denoted as $\hat{\boldsymbol{G}'}^k_{t|t-1}$ and $\hat{\boldsymbol{G}'}^k_t$ here, respectively. And the estimate error

covariance is denoted as $\boldsymbol{M}_{t|t-1}^k = Cov(\overline{\boldsymbol{G}}_t|\boldsymbol{U}_{[t-1]}, \boldsymbol{G}_{[t-1]})$ and $\boldsymbol{M}_t^k = Cov(\overline{\boldsymbol{G}}_t|\boldsymbol{U}_{[t]}, \boldsymbol{G}_{[t]})$. It is easy to calculated that

$$Cov(\overline{\boldsymbol{G}}_t|\boldsymbol{U}_{[t-1]}, \boldsymbol{G}_{[t]}) = (\frac{1}{\boldsymbol{M}_{t|t-1}^k} + \frac{1}{\boldsymbol{Q}_t^k})^{-1} \tag{58}$$

Next, we establish the Kalman filter with the innovation ν_t^k

$$\boldsymbol{\nu}_t^k = \begin{bmatrix} \boldsymbol{U}_t^k \\ \boldsymbol{G}_t \end{bmatrix} + \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} \hat{\boldsymbol{G}'}^k_{t|t-1} = \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} (\overline{\boldsymbol{G}}_t - \hat{\boldsymbol{G}'}^k_{t|t-1}) + \begin{bmatrix} \boldsymbol{R}_t^k \\ \boldsymbol{E}_t^k \end{bmatrix} \tag{59}$$

$$\Sigma_{\boldsymbol{\nu}_t^k} = \begin{bmatrix} (f_t^k)^2 & f_t^k \\ f_t^k & 1 \end{bmatrix} \boldsymbol{M}_{t|t-1}^k + \begin{bmatrix} \sigma^2_{\boldsymbol{R}_t^k} & 0 \\ 0 & \boldsymbol{Q}_t^k \end{bmatrix} \tag{60}$$

Then the filter form of the estimate is

$$\hat{\boldsymbol{G}'}^k_{t|t-1} = a'^k_{t-1} \hat{\boldsymbol{G}'}^k_{t-1|t-1} \tag{61}$$

$$\boldsymbol{M}_{t|t-1}^k = (a'^k_{t-1})^2 \boldsymbol{M}_{t-1|t-1}^k + \sigma^2_{\boldsymbol{W}'^k_{t-1}} \tag{62}$$

$$\hat{\boldsymbol{G}'}^k_t = \hat{\boldsymbol{G}'}^k_{t|t-1} + \boldsymbol{K}_t^k \boldsymbol{\nu}_t^k \tag{63}$$

$$\boldsymbol{M}_t^k = \left(\boldsymbol{I} - \boldsymbol{K}_t^k \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} \right) \boldsymbol{M}_{t|t-1}^k \tag{64}$$

$$= \left(\frac{(f_t^k)^2}{\sigma^2_{\boldsymbol{R}_t^k}} + \frac{1}{\boldsymbol{Q}_t^k} + \frac{1}{\boldsymbol{M}_{t|t-1}^k} \right)^{-1} \tag{65}$$

With the gain of Kalman Filter as $\boldsymbol{K}_t^k = \boldsymbol{M}_{t|t-1}^k \begin{bmatrix} f_t^k & 1 \end{bmatrix} \Sigma^{-1}_{\boldsymbol{\nu}_t^k}$. We need to relate the error of this estimate $\boldsymbol{M}_{t|t-1}^k$ and $\boldsymbol{M}_t^k$ with error $\boldsymbol{P}_{t|t-1}^k$ and $\boldsymbol{P}_t^k$ in above section. With the fact that

$$\boldsymbol{M}_t^k = Cov(\overline{\boldsymbol{G}}_t^k|\boldsymbol{U}_{[t]}^k, \boldsymbol{G}_{[t]}) = Cov(\overline{\boldsymbol{G}}_t^k - \boldsymbol{G}_t|\boldsymbol{U}_{[t]}^k, \boldsymbol{G}_{[t]}) \tag{66}$$

$$= Cov(\overline{\boldsymbol{G}}_t^k - \boldsymbol{G}_t|\boldsymbol{U}_{[t]}^k, \boldsymbol{G}_{[t]}, \hat{\boldsymbol{G}}_t^k) \tag{67}$$

$$= Cov(\overline{\boldsymbol{G}}_t^k - \boldsymbol{G}_t|\boldsymbol{G}_t - \hat{\boldsymbol{G}}_t^k) \tag{68}$$

We can let $\boldsymbol{G}_t - \overline{\boldsymbol{G}}_t^k \perp \overline{\boldsymbol{G}}_t^k - \hat{\boldsymbol{G}}_t^k$ because it is existed when choosing some parameters. Let $\hat{\boldsymbol{G}}_t^k = b_t^k \boldsymbol{U}_t^k + \boldsymbol{B}_t^k \boldsymbol{U}_{[t-1]}^k$ and $\overline{\boldsymbol{G}}_t^k = c_t^k \boldsymbol{G}_t^k + C_t^k \boldsymbol{G}_{[t-1]}^k$ with (54), the independent relationship is true when choosing $c_t^k = 1$, $b_t^k \boldsymbol{\Psi}_t^k - \boldsymbol{C}_t^k = b_t^k \boldsymbol{\Gamma}_t^k + \boldsymbol{B}_t^k = \boldsymbol{0}$ and $h_t^k b_t^k - c_t^k = 0$. Therefore

$$\boldsymbol{M}_{t|t}^k = Cov(\overline{\boldsymbol{G}}_t^k - \boldsymbol{G}_t|\boldsymbol{G}_t - \hat{\boldsymbol{G}}_t^k) = \boldsymbol{Q}_t^k (1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t}^k}) \tag{69}$$

Similarly, $\boldsymbol{M}_{t|t-1}^k$ can be related to $\boldsymbol{P}_{t|t-1}^k$ in the same way as

$$Cov(\overline{\boldsymbol{G}}_t^k|\boldsymbol{U}_{[t-1]}^k, \boldsymbol{G}_{[t]}) = \boldsymbol{Q}_t^k (1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_{t|t-1}^k}) \tag{70}$$

Putting (52) and (53) into (57) to finish the proof of this theorem. ∎

Combine the results in above would achieve Theorem 1 with the fact that function (18) is obtained directly from Theorem 3 and Theorem 4. The condition (19) is to guarantee the covariance in (70) greater than zero. The condition (20) is from (49) with the variance $\sigma^2_{\boldsymbol{V}_t^k}$ is greater than zero. And the last condition (21) is due to the constraint of distortion.

APPENDIX B
THEORETICAL ANALYSIS OF ALGORITHM 2

Firstly, we turn the optimal problem (P2) into the following optimal problem with the same optimal $\boldsymbol{P}_t^k$ to achieve rate-distortion function

$$(P3) \quad R'_t(D_t) = \inf_{\boldsymbol{P}_t^{[K]}} \frac{1}{2} \log(\sum_{k=1}^K \frac{1}{\boldsymbol{P}_t^k} - \frac{K-1}{\sigma^2_{\boldsymbol{G}_t^{des}}}) - \sum_{k=1}^K \frac{1}{2} \log(1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_t^k}) \tag{71}$$

$$s.t. \quad \boldsymbol{P}_t^k > \boldsymbol{Q}_t^k \tag{72}$$

$$\boldsymbol{P}_t^k \geq (\frac{1}{\sigma_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k})^{-1} \tag{73}$$

$$\sum_{k=1}^K (\frac{1}{\boldsymbol{P}_{t|t-1}^k} - \frac{1}{\sigma^2_{\boldsymbol{G}_t}}) + \frac{1}{\sigma^2_{\boldsymbol{G}_t}} \geq \frac{1}{D_t} \tag{74}$$

The function of $R'_t(D_t)$ is the convex of the factor $\boldsymbol{P}_t^{[K]}$. Then, we turn it into Lagrangian function

$$L(D_t) = \log \left[\sum_{j \neq k} (\frac{1}{\boldsymbol{P}_t^j} - \frac{1}{\boldsymbol{G}_t}) + \frac{1}{\boldsymbol{P}_t^k} \right] - \sum_{k=1}^K \log \left[1 - \frac{\boldsymbol{Q}_t^k}{\boldsymbol{P}_t^k} \right] \tag{75}$$

$$- \sum_{k=1}^K \boldsymbol{\lambda}_t^k \left[\boldsymbol{P}_t^k - \left(\frac{1}{\sigma^2_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k} \right)^{-1} \right] \tag{76}$$

$$- \boldsymbol{\nu}_t \left[\sum_{j \neq k} (\frac{1}{\boldsymbol{P}_t^j} - \frac{1}{\boldsymbol{G}_t}) + \frac{1}{\boldsymbol{P}_t^k} - \frac{1}{D_t} \right] \tag{77}$$

$$s.t. \quad \boldsymbol{P}_t^k > \boldsymbol{Q}_t^k \tag{78}$$

Assume that $\tilde{\boldsymbol{\Gamma}}_t^k = \sum_{j \neq k} (\frac{1}{\boldsymbol{P}_t^j} - \frac{1}{\boldsymbol{G}_t})$

The KKT condition is

$$\frac{\partial L(D_t)}{\partial \boldsymbol{P}_t^k} = 0 \tag{79}$$

$$\boldsymbol{\lambda}_t^k \left[\boldsymbol{P}_t^k - \left(\frac{1}{\sigma^2_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k} \right)^{-1} \right] = 0 \tag{80}$$

$$\boldsymbol{\nu}_t \left[\boldsymbol{\Gamma}_t^k + \frac{1}{\boldsymbol{P}_t^k} - \frac{1}{D_t} \right] = 0 \tag{81}$$

$$\boldsymbol{P}_t^k > \boldsymbol{Q}_t^k \tag{82}$$

(a) If $\boldsymbol{P}_t^k > \left(\frac{1}{\sigma^2_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}_{t|t-1}^k} \right)^{-1}$, $\lambda_t^k = 0$. Then from (79),

we can get that

$$\boldsymbol{\nu}_t = \frac{1}{\boldsymbol{\Gamma}_t^k + \frac{1}{\boldsymbol{P}_t^k}} + \frac{\boldsymbol{Q}_t^k}{1 - \boldsymbol{Q}_t^k \frac{1}{\boldsymbol{P}_t^k}} \tag{83}$$

$$\frac{1}{\boldsymbol{P}_t^k} = \frac{1}{2}\left[(\frac{1}{\boldsymbol{Q}_t^k} - \boldsymbol{\Gamma}_t^k) + \sqrt{(\frac{1}{\boldsymbol{Q}_t^k} + \boldsymbol{\Gamma}_t^k)^2 - \frac{4(\frac{1}{\boldsymbol{Q}_t^k} + \boldsymbol{\Gamma}_t^k)}{\boldsymbol{\nu}_t}}\right] \tag{84}$$

(b) If $\boldsymbol{P}_t^k = \left(\frac{1}{\sigma^2_{\boldsymbol{N}_t^k}} + \frac{1}{\boldsymbol{P}^k_{t|t-1}}\right)^{-1}$, $\boldsymbol{\lambda}_t^k = \frac{1}{(\boldsymbol{P}_t^k)^2}(\boldsymbol{\nu} - \frac{1}{\boldsymbol{\Gamma}_t^k + \frac{1}{\boldsymbol{P}_t^k}} - \frac{\boldsymbol{Q}_t^k}{1-\boldsymbol{Q}_t^k \frac{1}{\boldsymbol{P}_t^k}})$. Implying that $\boldsymbol{\nu} \geq \frac{1}{\boldsymbol{\Gamma}_t^k + \frac{1}{\boldsymbol{P}_t^k}} - \frac{\boldsymbol{Q}_t^k}{1-\boldsymbol{Q}_t^k \frac{1}{\boldsymbol{P}_t^k}}$
Then, the equations (25) and (26) are obtained.

REFERENCES

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

[2] O. A. Wahab, A. Mourad, H. Otrok, and T. Taleb, "Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems," *IEEE Commun. Surv. Tut.*, vol. 23, no. 2, pp. 1342–1397, 2021.

[3] D. Fan, X. Yuan, and Y.-J. A. Zhang, "Temporal-structure-assisted gradient aggregation for over-the-air federated edge learning," *IEEE J. Sele. Areas Commun.*, vol. 39, no. 12, pp. 3757–3771, 2021.

[4] N. Nguyen-Thanh, P. Ciblat, S. Maleki, and V.-T. Nguyen, "How many bits should be reported in quantized cooperative spectrum sensing?" *IEEE Wireless Commun. Lett.*, vol. 4, no. 5, pp. 465–468, 2015.

[5] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Sys*, 2017, pp. 1709–1720.

[6] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, "vqSGD: Vector quantized stochastic gradient descent," *IEEE Trans. Inf. Theory*, vol. 68, no. 7, pp. 4573–4587, 2022.

[7] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," in *in Proc. of IEEE ICASSP*, 2021, pp. 3110–3114.

[8] E. Rizk, S. Vlaski, and A. H. Sayed, "Dynamic federated learning," in *Proc. IEEE 21st Int. Workshop on Signal Proce. Advances in Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

[9] N. Zhang, M. Tao, and J. Wang, "Sum-rate-distortion function for indirect multiterminal source coding in federated learning," in *2021 Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2161–2166.

[10] J. Chen, X. Zhang, T. Berger, and S. Wicker, "An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the CEO problem," *IEEE J. Sele. Areas Commun.*, vol. 22, no. 6, pp. 977–987, 2004.

[11] V. Kostina and B. Hassibi, "The CEO problem with inter-block memory," *IEEE Trans. Inf. Theory*, vol. 67, no. 12, pp. 7752–7768, 2021.

[12] ——, "Rate-cost tradeoffs in scalar LQG control and tracking with side information," in *2018 Proc. Annual Allerton Conf. on Commun., Control, and Comput. (Allerton)*, 2018, pp. 421–428.

[13] G. Yan, S.-L. Huang, T. Lan, and L. Song, "DQ-SGD: Dynamic quantization in SGD for communication-efficient distributed learning," in *IEEE 18th Inter. Conf. on Mob. Ad Hoc and Smart Syst. (MASS)*, 2021, pp. 136–144.

[14] A. Abdi and F. Fekri, "Reducing communication overhead via CEO in distributed training," in *Proc. IEEE 20st Int. Workshop on Signal Proce. Advances in Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[15] Y. Huiyuan, D. Tian, and Y. Xiaojun, "Federated learning with lossy distributed source coding: Analysis and optimization." [Online]. Available: https://arxiv.org/abs/2204.10985