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 with limited throughput 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 central server reconstructs model parameters under distortion constraints. Furthermore, when considering the iterative process, we utilize the Kalman filter to reduce the computational complexity. And in each iteration, a generalized waterfilling algorithm is used to solve the optimal quantization level of the local client parameter dimension. Numerical results show that the rate-distortion optimization adaptive gradient quantization approach proposed in this paper reduces communication costs compared with conventional quantization methods.**

***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. However, the data often contains private information, that local clients may prefer not to share. To solve this issue, federated learning(FL) allows multiple clients to jointly train a machine learning model on their combined data, without revealing their data to a central server [1].

However, one of the major challenges of FL is that constantly updating model parameters can lead to a massive communication overload, since the parameters could be in the tens of millions for deep neural networks like ResNet [2]. This significantly slows down the convergence of FL because the communication link between local clients and the central server is typically constrained [3]. Especially for FL with over-the-air computation (AirComp) over multiple access channels, model parameters on local gradients should be transmitted to the central server under the bandwidth limitation [4].

Therefore, many researches effort on communication efficiency approaches to reduce the communication overhead caused by the message exchange between local clients and the central server. One of the most popular methods is quantization. Quantization involves lossy compression of gradient vectors through quantizing entry to a finite-bit low precision value [5]. [6] firstly applied stochastic uniform quantizer on gradient of model parameters. [7] followed this work and proposed an adaptive quantization strategy to generate approximate quantization levels for each round. The work in [8] considered a hierarchical gradient quantization method. Additional forms of probabilistic scalar quantization for FL were considered in [9].

These works significantly reduce the communication burden of FL systems at the cost of degrading model performance or increasing computational complexity. Furthermore, some researchers exlopited the drift of global model parameters in FL [10]. In this paper, we investigate communication efficient FL with a rate-distortion optimization adaptive gradient quantization(RDO-AGQ) approach using lossy source coding theory. The first attempt was made by Zhang et al., who introduced rate-distortion approach to FL in [11]. However, they only considered one iteration round and came to a very weak conclusion. Inspired by [11], we propose a computable algorithm for FL and improve traditional rate-distortion function during tracking the parameters in both local clients and the central server.

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 [12] using the "Generalized Waterfilling" approach. The work in [13] extended the Gaussian CEO problem by tracking and proposed a suboptimal waterfilling allocation algorithm at last. Inspired by Kalman filter utilized in causal source coding theory [14], we propose a novel RDO-AGQ approach to compress parameters before upload. The main contributions of this paper are summarized as follows:

- We propose a novel RDO-AGQ approach for communication efficient federated learning. Different from [7], our approach is able to calculate the adaptive quantization levels not only for each communication round but also for local client parameter gradients under different label

The first two authors make equal contributions to this work.

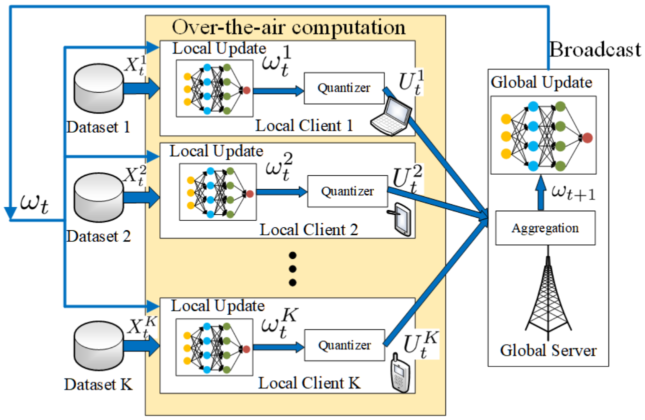

Fig. 1. Conventional federated learning framework

dimensions.

- In order to reduce the communication costs, we introduce the rate-distortion approach into FL framework. Compared with conventional FL framework, our rate-distortion optimization approach considers the drift of global model parameters and aims to minimize the communication costs under distortion constraints. Different from [11], we consider the memory of local clients and central server and study the tracking problem to obtain optimal quantization strategy.
- Numerical results validate the effectiveness of the proposed approach and show that this approach significantly reduces the upload communication cost when comparing with conventional FL system using quantization approach [6], [7].

The rest of this paper is organized as follows. Section II presents the preliminaries about the conventional FL system. Section III details the framework of the proposed RDO-AGQ. Theoretical analysis of how the proposed approach works is present 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 $(X^1, X^2, \ldots, X^K)$ in clients dimension. And $X_{[t]}$ denotes the random vector $(X_1, X_2, \ldots, X_t)$ in iterative rounds dimension.

II. Preliminary of Federated Learning

In this section, we present necessary preliminaries about conventional FL framework. As depicted in Fig.1, we consider a conventional FL framework, which consists of K local clients and one central global server dispersed in space. Training data, distributed among local clients, is either IID or non-IID distributed. Let $\{X_t^k\}_{i=1}^{n_k}$ be the set of n_k labeled training samples available at the kth client, $k \in \{1, 2, \ldots, K\}$.

Each client has access to local dataset $\{X_t^k\}_{i=1}^{n_k}$ and the global learning model ω_t. To minimize the objective function $F^k(\cdot)$, the kth client trains a machine learning model, represented by the parameter set ω_t^k. These local objective functions are defined as the empirical average over the corresponding training set i.e.,

$$F_t^k(\omega_t; \{X_t^k\}_{i=1}^{n_k}) \triangleq \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_t^k(\omega_t; X_t^k), \quad (1)$$

where $\ell_t^k(\cdot;\cdot)$ is the loss function. After the local training phase, all clients transmit updated parameters to the global server. The communication cost occurring during this phrase can be prohibitive since every client in the system needs to send all model parameters to the global server [15]. Based on methods used in numerous researches, e.g. Stochastic Gradient Descent(SGD) [16], quantization [6], [7], [9], Sparsification [17]. RDO-AGQ is exploited in this paper as a new perspective to generate the optimal quantization levels and compress parameter gradients using stochastic uniform quantizers, which will be explained in detail in the next section.

After receiving the entire set of K compressed parameters $\{U_t^1, \ldots, U_t^K\}$, which is the quantized form of desired local model parameters $\{\omega_t^1, \ldots, \omega_t^K\}$ from clients, the central server aggregates these parameters in a certain way to generate one set of updated parameters. Federated averaging(FA) algorithm is employed here, which is a wildly used method to aggregates parameters [15], [18]. At last, a new training iteration is started by broadcasting the global updated set of model parameters to all local clients.

The whole FL model aims at recovering model parameters ω^o by seeking the minimization of the average risk satisfying

$$\omega^o \triangleq \arg \min_{\omega_t^1, \ldots, \omega_t^K} \left\{ F_t(\omega_t) = \sum_{k=1}^{K} \frac{1}{K} F_t^k(\omega_t; \{X_t^k\}_{i=1}^{n_k}) \right\}. \quad (2)$$

Communication between local clients and the global server occurs during parameter upload and download phases. In general, the communication cost of the latter is negligible since the transmission of the global parameters to all clients can be done in a broadcast manner [18]. However, during the upload phase, each client has to unicast a diverse set of parameters to the central server, which can incur prohibitive communication costs. Accordingly, we concentrate on minimizing the communication cost during upload phase in this paper.

III. The Framework of RDO-AGQ for FL

This section presents the RDO-AGQ framework for FL system. Gradients are transmitted between local clients and a central server to make the communication more private and efficient. Before compressing the gradients using stochastic uniform quantizers, we introduce the causal multiterminal source coding method into FL to determine the adaptive quantized levels for each label dimension gradient. At last, a rate-distortion optimization problem is proposed under some assumptions.

A. Federated Learning Framework

We consider an FL system where each local client uses SGD to train a machine learning model. After updating local parameters ω_t^k at each client, the recursion (1) becomes:

$$\omega_t^k = \omega_t - \eta \nabla_{\omega_t^k} F_t^k(\omega_t; \{X_t^k\}_{i=1}^{n_k}). \quad (3)$$

In order to protect the privacy of each client, the kth client only transmits the gradient, which is defined as

$$G_t^k \triangleq \nabla_{\omega_t^k} F_t^k(\omega_t; \{X_t^k\}_{i=1}^{n_k}), \tag{4}$$

to the central server in local training phase. According to (2), one widely used method for aggregating received parameters by the central server is FA, which we use as a baseline in this paper. The global updated gradient aggregated by FA is called desired gradient, denoted as G_t^{des}, via:

$$G_t^{des} = \frac{1}{K}\sum_{k=1}^{K} G_t^k. \tag{5}$$

Naturally, the global model parameters updated by desired gradient is known as desired model parameters, represented as ω_t^{des}, via $\omega_{t+1}^{des} = \omega_t^{des} - \eta G_t^{des}$.

Since upload throughput is typically more limited than its download counterpart [19], the kth client needs to transmit a finite-bit compressed representation of its model updates. Then, for efficient communication, we propose RDO-AGQ, which will be describe in detail in the following section.

B. Framework of RDO-AGQ

In this section, we propose a novel rate-distortion-based adaptive quantization approach, which conveys the local model update gradients $\{G_t^k\}_{k=1}^K$ from local clients to the central server via an ideal bit-constrained channel. In the absence of computational constraints and storage limits, this method consists of K quantizers with memory, which means quantizers and central server are able to use data in the pervious iterative rounds.

For the tth iteration, the kth model update gradient $\{G_t^k\}$ is quantized as R_t^k bits, denoted as $U_t^k \in \{0, 1, \ldots, 2^{R_t^k} - 1\} \triangleq \mathcal{U}_t^K$. Modeling the uplink channel as a bit-constrained link is a standard assumption in the FL literature [9], [16], [17], [20]–[26]. However, aiming at reducing the communication costs without incurring large losses, the center server aggregates received gradients under distortion constraints in RDO-AGQ FL system. In other words, the estimate $\hat{G}_t$, reconstructed by the global sever using the received compressed parameters $\{U_t^{[K]}\}$, allows an error with desired average gradient $\hat{G}_t^{des}$ under distortion constraint D_t. Here, $\hat{G}_t$ is an estimate of the desired average gradient G_t^{des}. The average distortion between estimated average gradient $\hat{G}_t$ and the desired average gradient G_t^{des} need to be less than a given distortion constraint D_t, via:

$$\mathbb{E}[d(G_t^{des}, \hat{G}_t)] \leq D_t. \tag{6}$$

At last the centralized gradient descent step takes the form

$$\omega_{t+1} = \omega_t - \eta \hat{G}_t \tag{7}$$

to update the model parameters in the central server. And then the central server broadcasts the newly updated global model ω_{t+1} to every client to complete one model update process.

In the presence of quantization, distortion may degrade the ability of the central server to update its model. Choosing an appropriate distortion constraint would not greatly affect the accuracy of the FL model. However, due to uplink communication resources are precious, reducing communication costs, e.g. communication rates for total iterative rounds, are considered via RDO-AGQ.

C. Problem Formulation

The goal of RDO-AGQ is to find the optimal quantization levels to minimize communications costs for all uplink channels, denoted as $R_t \triangleq \sum_{k=1}^K R_t^k$. 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 minimizer ω_t^{des}, namely that it follows a random walk model as in [10].

Assumption 1: We assume that the desired model ω_t^{des} follows a random walk:

$$\omega_t^{des} = a * \omega_{t-1}^{des} + W_{t-1} \tag{8}$$

where W_t denotes some zero mean random variable independent of $\omega_{t'}^{des}$ for any $t' < t$ and with bounded variance, i.e., $\mathbb{E}[||W_t||^2] = \sigma_{W_t}^2$.

Then, we assume the gradient follows a typical distribution for theorem analysis, which is proved to be approved in [11].

Assumption 2: The both local and global gradient, namely G_t^k and G_t^{des} are Gaussian random variables with zero mean.

Moreover, we assume the distortion between estimated average gradient $\hat{G}_t$ and desired average gradient G_t^{des} is under a specifical measure.

Assumption 3: The distortion function is under squared error distortion measure for Gaussian random variables:

$$d(G_t^{des}, \hat{G}_t) = ||G_t^{des} - \hat{G}_t||^2. \tag{9}$$

At last, we assume that only local clients know the random seed and mitigate the effects of perturbation through our coding strategy which is easy to implement in reality than [19].

Assumption 4: We assume the change in local gradient G_t^k across clients by adding randomly sampled constants $C_t^k \sim \mathcal{N}_t^k(0, \sigma_{C_t^k}^2)$, i.e.,

$$\tilde{G}_t^k = G_t^k + C_t^k. \tag{10}$$

In light of the above assumptions, the desired global updated gradient G_t^{des} follows the same random walk model as desired global updated model ω_t^{des}, via

$$G_t^{des} = a * G_{t-1}^{des} + W_{t-1}. \tag{11}$$

With Assumption 2, Assumption 4 and the property of Gaussian variables, it is sensible to let

$$\tilde{G}_t^k = G_t^{des} + N_t^k. \tag{12}$$

Where, $\{W_t, N_t^1, \ldots, N_t^K\}_{t=1}^T$ are Gaussian random variables independent of G_t^{des} with independent components; each component of W_t is distributed as $\mathcal{N}(0, \sigma_{W_t}^2)$, and each component of N_t^k is distributed as $\mathcal{N}(0, \sigma_{N_k}^2)$.

At last, because our RDO-AGQ aims to calculate the minimum communication rate under given distortion contractions. We propose the problem formulation below.

$$(P1) \quad R_t(D_t) = \min_{U_t^K} \sum_{k=1}^{K} R_t^k \tag{13}$$

$$s.t. \quad \mathbb{E}[||G_t^{des} - \hat{G}_t||^2] \leq D_t. \tag{14}$$

Here, the constraint (14) is under Assumption 3. And $\hat{G}_t$ is the estimated gradient of G_t^{des}, which follows the optimal quantization levels to achieving the minimum expected squared error:

$$\hat{G}_t \triangleq \mathbb{E}\left[G_t^{des}|U_{[t]}^{[K]}\right] \tag{15}$$

In light of above assumptions and problem formulation, we propose a solution of problem P1 in next section.

IV. RDO-AGQ Algorithm

In this section, we investigate minimum total communication rate in the upload phase and propose an algorithm to calculate the quantization levels for each label dimension gradient. Specifically, this scheme utilizes the memory of local updated model gradients with Kalman filters to predict the local updated model gradients in the current iteration round. Both the computational complexity and the communication costs are reduced due to the utilization of Kalman filter. A iterative water-filling algorithm for calculating the communication rate is also introduced here to solve quantization levels. Finally, combined with the above derivation, the RDO-AGQ algorithm is proposed.

A. Problem Conversion with Kalman Filter

Before conversing problem P1 into a computable form, let us make some notations here. In the central server, we define the variance of desired global updated model gradient G_t^{des} as $\sigma^2_{G_t^{des}}$. Let the quantized updated model gradient of kth local client be $\hat{G}_t^k$, via:

$$\hat{G}_t^k = \mathbb{E}[G_t^{des}|U_{[t]}^k]. \tag{16}$$

And the predicted updated model gradient for kth local client, using past $t-1$ iterative rounds data, be $\hat{G}_{t|t-1}^k$, via:

$$\hat{G}_{t|t-1}^k = \mathbb{E}[G_t^{des}|U_{[t-1]}^k]. \tag{17}$$

Moreover, let the error variance between the desired global updated model gradient G_t^{des} and the received updated model gradient $\hat{G}_t^k$ be P_t^k, via:

$$P_t^k = \mathbb{E}[(G_t^{des} - \hat{G}_t^k)^2]. \tag{18}$$

And let the error variance between desired global updated model gradient G_t^{des} and the predicted updated model gradient $\hat{G}_{t|t-1}^k$ be $P_{t|t-1}^k$, via:

$$P_{t|t-1}^k = \mathbb{E}[(G_t^{des} - \hat{G}_{t|t-1}^k)^2]. \tag{19}$$

Local clients also need to estimate their updated model gradients when observing the perturbed gradient $\tilde{G}_t^k$ before quantization. Let the updated model gradient estimated by the kth local client $\tilde{G}_t^k$ be $\overline{G}_t^k$. Similarly, the updated model gradient predicted by the kth local client using data from the past $t-1$ iteration rounds is $\overline{G}_{t|t-1}^k$. Naturally, we define the error variance between the desired global updated model gradient G_t^{des} and the received or predicted updated model gradient $\overline{G}_t^k$ or $\overline{G}_{t|t-1}^k$ be Q_t^k and $Q_{t|t-1}^k$, respectively, via:

$$Q_t^k = \mathbb{E}[(G_t^{des} - \overline{G}_t^k)^2] \tag{20}$$

$$Q_{t|t-1}^k = \mathbb{E}[(G_t^{des} - \overline{G}_{t|t-1}^k)^2] \tag{21}$$

Using the above definitions, 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 derived as the following optimal problem.

$$(P2) \quad R_t(D_t) = \inf_{P_{t|t}^{[K]}} \frac{1}{2} \log \frac{\sum_{k=1}^{K} \frac{1}{P_{t|t}^k} - \frac{K-1}{\sigma^2_{G_t^{des}}}}{\sum_{k=1}^{K} \frac{1}{P_{t|t-1}^k} - \frac{K-1}{\sigma^2_{G_t^{des}}}} \tag{22}$$

$$+ \sum_{k=1}^{K} \frac{1}{2} \log \frac{(1 - \frac{Q_t^k}{P_{t|t-1}^k})}{(1 - \frac{Q_t^k}{P_{t|t}^k})}.$$

$$s.t. \quad P_{t|t}^k > Q_t^k \tag{23}$$

$$P_{t|t}^k \geq (\frac{1}{\sigma^2_{N_t^k}} + \frac{1}{P_{t|t-1}^k})^{-1} \tag{24}$$

$$\sum_{k=1}^{K} (\frac{1}{P_{t|t-1}^k} - \frac{1}{\sigma^2_{G_t^{des}}}) + \frac{1}{\sigma^2_{G_t^{des}}} \geq \frac{1}{D_t} \tag{25}$$

with $P_{t|t-1}^k$ and P_t^k satisfying

$$P_{t|t-1}^k = a_{t-1}^2 P_{t-1} + \sigma^2_{W_t}, \tag{26}$$

$$P_t^k = (\frac{1}{\sigma^2_{N_t^k} + \frac{\sigma^2_{V_t^k}}{(h_t^k)^2}} + \frac{1}{P_{t|t-1}^k})^{-1}. \tag{27}$$

Where, D_t is the given distortion constraint and $\sigma^2_{N_t^k}$ is the variance of independent Gaussian random variable N_t^K in (12).

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

$$R_t^k \geq I(\tilde{G}_{[t]}^{[K]}; U_t^{[K]}|U_{[t-1]}^{[K]}) \tag{28}$$

Afterwards, Kalman filter would be utilized for estimating global updated model gradient G_t^{des} using observing data before quantizers. Jointing the Kalman filter result with following Lemma 1, this theorem is easy to proved by simple operation.

Lemma 1 (Theorem 4 of [13]): For all $\sigma^2_{X_T|U_{[T]}^{[K]}} < D_T < \sigma^2_X$, the causal CEO rate-distortion function for the Gauss-Markov source is given by

$$R_{\sum_T}(D_T) = \frac{1}{2} \log \frac{\tilde{D}_T}{D_T} + \min_{\{d_T^k\}_{k=1}^K} \sum_{k=1}^{K} \frac{1}{2} \frac{\tilde{d}_T^k - \sigma^2_{X_T|U_{[T]}^{[K]}}}{d_T^k - \sigma^2_{X_T|U_{[T]}^{[K]}}} \frac{d_T^k}{\tilde{d}_T^k}, \tag{29}$$

where

$$X_T = aX_{t-1} + V \tag{30}$$

$$\tilde{D}_T \triangleq a^2 D_T + \sigma_V^2 \tag{31}$$

$$\tilde{d}_T^k \triangleq a^2 d_T^k + \sigma_V^2, \tag{32}$$

and the minimum is over d_T^k, $k \in [K]$, that satisfy

$$\frac{1}{d} \leq \frac{1}{\sigma^2_{X_T|U_{[T]}^{[K]}}} - \sum_{k=1}^{K} \left(\frac{1}{\sigma^2_{X_T|U_{[T]}^{[K]}}} - \frac{1}{d_T^k} \right), \tag{33}$$

$$\sigma^2_{X_T|U_{[T]}^{[K]}} \leq d_T^k \leq \sigma_X^2 \tag{34}$$

■

B. Iterative Waterfilling Algorithm

In this section, we propose an iterative water-filling algorithm to calculate the numerical value of the above optimal problem (P2) in Theorem 1 at each iterative round t. Because the error covariance $P_{t|t-1}^k$ is calculated by the time slot $t-1$, this time we only need to find the optimal $P_{t|t}^k$. To calculate the minimum communication rate is equivalent to finding the optimal error covariance of estimate $P_{t|t}^k$.

Algorithm 1 Iterative Waterfilling Algorithm

1: **Initialization**: $P_{t|t}^k$ 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 $\Gamma_t^k = \sum_{j \neq k} (\frac{1}{P_{t|t}^j} - \frac{1}{\sigma^2_{G_t^{des}}})$.
5:     Update objective parameter $P_{t|t}^k$ with (35)
6:     Update communication rate threshold ν_t with (36)
7:   **end for**
8:   Update $R_t(D_t)$ with (22)
9: **end while**

where

$$\frac{1}{P_{t|t}^k} = \frac{1}{2} \left[(\frac{1}{Q_t^k} - \Gamma_t^k) + \sqrt{(\frac{1}{Q_t^k} + \Gamma_t^k)^2 - \frac{4(\frac{1}{Q_t^k} + \Gamma_t^k)}{\nu_t}} \right] \tag{35}$$

$$\nu_t = \arg \left\{ \sum_{k=1}^{K} (\frac{1}{P_{t|t}^k} - \frac{1}{\sigma^2_{G_t^{des}}}) + \frac{1}{\sigma^2_{G_t^{des}}} = \frac{1}{D_t} \right\} \tag{36}$$

The quantization levels assignment algorithm is presented in Algorithm 1. From this algorithm, the total communication rate of the current iteration round is the output value $R_t(D_t)$ in line 8. And the quantization levels for each parameter label dimension are easy to operate by threshold ν_t. This algorithm is an expended form of iterative water-filling algorithm. And the theoretical analysis of each updating step in line 4 to line 6 is proposed in Appendix B.

C. RDO-AGQ Algorithm

We elaborate here how to quantify and transfer parameter gradients in a communication efficient FL setting. The detailed implementation of RDO-AGQ is shown in Algorithm 2.

Algorithm 2 Algorithm of RDO-AGQ FL

Input: Total number of clients K; Total number of communication rounds T; Local dataset $(X^1, X^2, \ldots, X^K)$; The set of distortion constrain $(D_1, D_2, \ldots, D_T)$;
Initialization: Global model parameter ω_0;
for $t = 1$ to T **do**
  for $k = 1$ to K **do**
    Calculate $G_t^k = \nabla_{\omega_t^k} F_t^k(\omega_t; \{X_t^k\}_{i=1}^{n_k})$
    Update local model: $\omega_t^k = \omega_{t-1}^K - \eta G_t^k$
    Quantize G_t^k with quantization levels with (28)
  end for
  Reconstructs G_t with FA.
  Update global model as $\omega_{t+1} = \omega_t - \eta \hat{G}_t$.
  Broadcast ω_{t+1} to every client.
end for

Through the pseudo code, it is intuitive to comprehend our RDO-AGQ algorithm. Firstly, during the local training phase, each local client uses SGD to train a machine learning model. Naturally, they need to update their own models. Before uploading, enter the quantification stage. We use the iterative waterfilling algorithm, presented in Algorithm 1, to achieve the sum communication rate of each iterative round and the quantization levels for each local client. During upload phase, each local client uploads their quantized parameters to the central server through ideal MAC. At last, the central server aggregates the global model in global update phase and broadcast the newly updated global model to all clients in download phase.

V. Numerical Results

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

A. 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 MNIST 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.

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

- Total amount of clients: K = 10.
- Gradient descent algorithm: SGD [16].
- Quantizer: Stochastic uniform quantizer [6], [7].
- Local update uses Vanilla CNN architecture with 10 epochs in each iterative round.

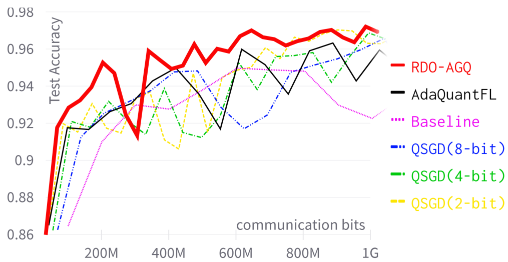

Fig. 2. Test accuracy versus communication costs. IID partition on MNIST.

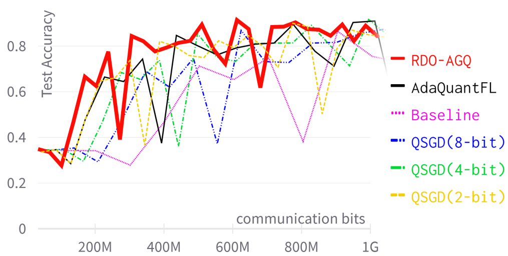

Fig. 3. Test accuracy versus communication costs. Non-IID partition on MNIST.

B. *Performance of RDO-AGQ*

In this section, we consider the performance of proposed RDO-AGQ approach. In general, RDO-AGQ only needs fewer total communication bits to achieve the specific train accuracy. And with the same communication costs, the FL system is able to achieve higher train accuracy using RDO-AGQ. Our experimental results verify that the proposed RDO-AGQ is able to achieve high train accuracy using fewer communication bits in most cases. As seen in Fig. 2, RDO-AGQ achieves the highest test accuracy with same communication costs when the dataset is IID distributed. In detail, the test accuracy only demands 340M bits to reach 96%, while the second-least communication costs to reach 96% test accuracy is obtained by AdaQuantFL, which consumes approximately 610M bits. Moreover, with consuming the same communication cost-s(e.g. about 600M bits), RDO-AGQ brings a test accuracy of 96.7%, while the test accuracy is only 95.8% and 95.2% for AdaQuantFL and 4-bit QSGD respectively.

When training non-IID MNIST dataset, observing Fig. 3, the proposed RDO-AGQ is still the best quantization approach. RDO-AGQ only costs approximately 510M bits to achieve 90% test accuracy, which consumes the fewest communication bits among all compared quantization methods. While the AdaQuantFL approach and 4-bit QSGD require approximately 775M bits and 828M bits, respectively, to achieve the same test accuracy. Moreover, the proposed RDO-AGQ also achieves the highest test accuracy in most situations with the same communication bits among all comparison quantization approach.

In conclusion, compared to AdaQuantFL and QSGD with d-ifferent fixed quantization levels, the performance of proposed RDO-AGQ approach is the best for MNIST in FL system. Our RDO-AGQ approach is able to reduce the communication costs in FL systems and shows robustness during simulation.

VI. Conclusion

This paper introduces an RDO-AGQ approach for communication-efficient federated learning. 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 proposed RDO-AGQ outperforms AdaQuanFL and QSGD for MNIST dataset with both iid and non-iid distributed.

VII. Appendix

A. *Proof of Theorem 1*

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

Lemma 2: [13, Appendix D] We let Gaussian random variable $X \sim \mathcal{N}(0, \sigma_X^2)$ with

$$Y_k = X + W_k, \quad k = 1, \ldots, K, \tag{37}$$

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

$$\mathbb{E}[X|Y_{[K]}] = \sum_{k=1}^{K} \frac{\sigma_{X|Y_{[K]}}^2}{\sigma_{W_k}^2} Y_k, \tag{38}$$

$$\frac{1}{\sigma_{X|Y_{[K]}}^2} = \frac{1}{\sigma_X^2} + \sum_{k=1}^{K} \frac{1}{\sigma_{W_k}^2}. \tag{39}$$

Proof: The proof of this lemma is detailed in [13, Appendix D]. ■

Then, using mutual information, problem P1 in (13) comes to the computation of rate-distortion function

$$R_t(D_t) = \inf I(\tilde{G}_{[t]}^{[K]}; U_t^{[K]} | U_{[t-1]}^{[K]}) \tag{40}$$

$$\overset{(a)}{=} \inf I(\overline{G}_{[t]}^{[K]}; U_t^{[K]} | U_{[t-1]}^{[K]}) \tag{41}$$

$$\overset{(b)}{=} \inf I((\overline{G}_{[t]}^{[K]}, G_{[t]}^{des}); U_t^{[K]} | U_{[t-1]}^{[K]}) \tag{42}$$

$$\overset{(c)}{=} \inf I(G_{[t]}^{des}; U_t^{[K]} | U_{[t-1]}^{[K]}) \tag{43}$$

$$+ \sum_{k=1}^{K} I(\overline{[t]}^{k}; U_t^k | U_{[t-1]}^k, G_{[t]}^{des}) \tag{44}$$

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

Afterwards, Kalman filter is used to estimate the global updated model gradient G_t in iterative round t by the each kth local client with $\tilde{G}_{[t]}^k$. Next theorem is the result of calculating the estimate of desired global updated model gradient G_t^{des} by local clients before coding.

Theorem 2: The iterative estimate of desired global updated model gradient G_t^{des} form is

$$\overline{G}_t^k = \frac{a * \sigma_{N_t^k}^2}{Q_{t|t-1}^k + \sigma_{N_t^k}^2}\overline{G}_{t-1}^k + \frac{Q_{t|t-1}^k}{Q_{t|t-1}^k + \sigma_{N_t^k}^2}\tilde{G}_t^k \tag{45}$$

$$= a'^k_{t-1}\overline{G}_{t-1}^k + W'^k_{t-1} \tag{46}$$

which follows:

$$Q_t^k = (\frac{1}{Q_{t|t-1}^k} + \frac{1}{\sigma_{N_t^k}^2})^{-1} \tag{47}$$

$$Q_{t|t-1}^k = a^2 * Q_{t-1|t-1}^k + \sigma_{N_t^k}^2 \tag{48}$$

Proof: This proof only need to use the standard Kalman Filer method. Let the $\mu_{\overline{G}_t}^k$ be the innovation and $K_{\overline{G}_t}^k$ be the gain of Kalman Filter. Therefore

$$\overline{G}_{t|t-1}^k = a * \overline{G}_{t-1} \tag{49}$$

$$Q_{t|t-1}^k = a^2 Q_{t-1}^k + \sigma_{W_{t-1}}^2 \tag{50}$$

$$\mu_{\overline{G}_t}^k = \tilde{G}_t^k - \overline{G}_{t|t-1}^k = \overline{G}_{t-1}^k - \overline{G}_{t|t-1}^k + N_t^k \tag{51}$$

$$K_{\overline{G}_t}^k = Q_{t|t-1}^k(Q_{t|t-1}^k + \sigma_{N_t^k}^2)^{-1} \tag{52}$$

$$\overline{G}_t^k = \overline{G}_{t|t-1}^k + K_{\overline{G}_t}^k \mu_{\overline{G}_t}^k \tag{53}$$

$$Q_t^k = (1 - K_{\overline{G}_t}^k)Q_{t|t-1}^k \tag{54}$$

Then, this theorem can be proved with some simple operation. ■

Next, we focus on calculating (43). The difference of this function with the standard Kalman Filter form is that it has K layers of $U_t^{[K]}$, which is seen as the observers. According to the problem formulation in Section III-C, let

$$U_t^k = h_t^k \tilde{G}_t^k + \Gamma_t^k U_{[t-1]}^k + \Psi_t^k \tilde{G}_{[t-1]}^k + V_t^k \tag{55}$$

$$= h_t^k G_t^{des} + \Gamma_t^k U_{[t-1]}^k + \Psi_t^k G_{[t-1]}^k + Z_t^k, \tag{56}$$

where h_t^k is a auxiliary factor, V_t^k is a Gaussian random variable independent of $(\tilde{G}_t^k,\ U_{[t-1]}^k,\ V_t^j)$, $j \neq k$ and $Z_t^k = V_t^k + h_t^k N_t^k + \Psi_t^k N_{[t-1]}^k$ is also Gaussian random variable independent of $(\tilde{G}_t^k,\ U_{[t-1]}^k,\ Z_t^j)$, $j \neq k$. It is obvious to find the value of Γ_t^k and Ψ_t^k has no effect of our objective $I(G_{[t]}^{des}; U_t^{[K]}|U_{[t-1]}^{[K]})$. So, let $\Gamma_t^k = \Psi_t^k = 0$ to get

$$U_t^k = h_t^k G_t^{des} + Z_t^k. \tag{57}$$

where, $Z_t^k = V_t^k + h_t^k N_t^k$ now. We estimate the desired global updated model gradient G_t^{des} by each local encoded codeword using Kalman Filter. The filter form of the estimate is

$$\hat{G}_{t|t-1}^k = a * \hat{G}_{t-1} \tag{58}$$

$$P_{t|t-1}^k = a^2 * P_{t-1} + \sigma_{W_t}^2 \tag{59}$$

$$K_{\hat{G}_t}^k = h_t^k P_{t|t-1}^k((h_t^k)^2 P_{t|t-1}^k + \sigma_{Z_t^k}^2)^{-1} \tag{60}$$

$$\hat{G}_t^k = \hat{G}_{t|t-1}^k + K_{\hat{G}_t}^k \mu_t^k \tag{61}$$

$$P_t^k = (1 - h_t^k K_{\hat{G}_t}^k)P_{t|t-1}^k \tag{62}$$

$$= (\frac{(h_t^k)^2}{\sigma_{Z_t^k}{}^2} + \frac{1}{P_{t|t-1}^k})^{-1} \tag{63}$$

The iterative form of (43) is shown in the following theorem.

Theorem 3: The iterative form of $I(G_{[t]}^{des}; U_t^{[K]}|U_{[t-1]}^{[K]})$ is

$$I(G_{[t]}^{des}; U_t^{[K]}|U_{[t-1]}^{[K]}) = \frac{1}{2}\log\frac{\sum_{k=1}^K \frac{1}{P_{t|t}^k} - \frac{K-1}{\sigma_{G_t^{des}}^2}}{\sum_{k=1}^K \frac{1}{P_{t|t-1}^k} - \frac{K-1}{\sigma_{G_t^{des}}^2}} \tag{64}$$

Proof: Firstly, we extend (43) as

$$I(G_{[t]}^{des}; U_t^{[K]}|U_{[t-1]}^{[K]}) = \frac{1}{2}\log\frac{Cov(G_t^{des}|U_{[t-1]}^{[K]})}{Cov(G_t^{des}|U_{[t]}^{[K]})} \tag{65}$$

using the characterization of mutual information for Gaussian random variables, where the function $Cov(\cdot)$ is covariance matrix. According to Lemma 2, the covariance in (65) is

$$Cov^{-1}(G_t^{des}|U_{[t-1]}^{[K]}) = \sum_{k=1}^K \frac{1}{P_{t|t-1}^k} - \frac{K-1}{\sigma_{G_t^{des}}^2} \tag{66}$$

$$Cov^{-1}(G_t^{des}|U_{[t]}^{[K]}) = \sum_{k=1}^K \frac{1}{P_{t|t}^k} - \frac{K-1}{\sigma_{G_t^{des}}^2} \tag{67}$$

This theorem is proved putting (66) and (67) into (65). ■

At last, we are interested in calculating (44) using the parameters of P_t^k and $P_{t|t-1}^k$. 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

$$U_t^k = h_t^k \tilde{G}_t^k + \Gamma_t^k U_{[t-1]}^k + \Psi_t^k X_{[t-1]}^k + V_t^k \tag{68}$$

$$= f_t^k \overline{G}_t^k + \Gamma_t^k U_{[t-1]}^k + \Phi_t^k G_{[t]}^{des} + R_t^k, \tag{69}$$

where the $1\times t$ vector $\Phi_t^k = [\Psi_t^k, h_t^k] - f_t^k \cdot \Sigma_{G_t^{des}, \tilde{G}_{[t]}^k} \cdot \Sigma_{\tilde{G}_{[t]}^k, \tilde{G}_{[t]}^k}^{-1}$, $R_t^k = V_t^k + [\Psi_t^k, h_t^k] \cdot N_{[t]}^k$ and f_t^k is the auxiliary variable. While the value of Φ_t^k and Γ_t^k makes no effect on $I(\overline{G}_t^k; U_t^k|U_{[t-1]}^k, G_{[t]}^{des})$. Therefore, we let $\Phi_t^k = \Gamma_t^k = 0$ to get $U_t^k = f_t^k \overline{G}_t^k + R_t^k$ and $G_t^{des} = \overline{G}_t^k + E_t^k$, where $\sigma_{E_t^k}^2 = Q_t^k$. Moreover, $R_t^k = Z_t^k = V_t^k + h_t^k \cdot N_t^k$ now. The iterative form of (44) is shown in the following theorem.

Theorem 4: The iterative form of $I(G_{[t]}^{des}; U_t^{[K]}|U_{[t-1]}^{[K]})$ is

$$I(\overline{G}_{[t]}^k; U_t^k|U_{[t-1]}^k, G_{[t]}^{des}) = \frac{1}{2}\log\frac{(1 - \frac{Q_t^k}{P_{t|t-1}^k})}{(1 - \frac{Q_t^k}{P_{t|t}^k})} \tag{70}$$

Proof: Firstly, we extend (44) as

$$I(\overline{G}_{[t]}^k; U_t^k|U_{[t-1]}^k, G_{[t]}^{des}) = \frac{1}{2}\log\frac{Cov(\overline{G}_t^k|U_{[t-1]}^k, G_{[t]}^{des})}{Cov(\overline{G}_t^k|U_{[t]}^k, G_{[t]}^{des})} \tag{71}$$

One of the nontrivial point here is that for the standard Kalman Filter method, it can only deal with the variables as $\mathbb{E}\{\overline{G}_t^k|U_{[t-1]}^k, G_{[t-1]}^{des}\}$ and $\mathbb{E}\{\overline{G}_t^k|U_{[t]}^k, G_{[t]}^{des}\}$ which is denoted as $\hat{G'}_{t|t-1}^k$ and $\hat{G'}_{t|t}^k$ here, respectively. And the estimate error covariance is denoted as $M_{t|t-1}^k = Cov(\overline{G}_t|U_{[t-1]}, G_{[t-1]}^{des})$ and $M_{t|t}^k = Cov(\overline{G}_t|U_{[t]}, G_{[t]}^{des})$. It is easy to calculated that

$$Cov(\overline{G}_t|U_{[t-1]}, G_{[t]}) = (\frac{1}{M_{t|t-1}^k} + \frac{1}{Q_t^k})^{-1} \tag{72}$$

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

$$\nu_t^k = \begin{bmatrix} U_t^k \\ G_t^{des} \end{bmatrix} + \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} \hat{G'}_{t|t-1}^k \tag{73}$$

$$= \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} (\overline{G}_t - \hat{G'}_{t|t-1}^k) + \begin{bmatrix} R_t^k \\ E_t^k \end{bmatrix} \tag{74}$$

$$\Sigma_{\nu_t^k} = \begin{bmatrix} (f_t^k)^2 & f_t^k \\ f_t^k & 1 \end{bmatrix} M_{t|t-1}^k + \begin{bmatrix} \sigma_{R_t^k}^2 & 0 \\ 0 & Q_t^k \end{bmatrix} \tag{75}$$

Then the filter form of the estimate is

$$\hat{G'}_{t|t-1}^k = a_{t-1}^{'k} \hat{G'}_{t-1|t-1}^k \tag{76}$$

$$M_{t|t-1}^k = (a_{t-1}^{'k})^2 M_{t-1|t-1}^k + \sigma_{W_{t-1}^{'k}}^2 \tag{77}$$

$$K_t^k = M_{t|t-1}^k \begin{bmatrix} f_t^k & 1 \end{bmatrix} \Sigma_{\nu_t^k}^{-1} \tag{78}$$

$$\hat{G'}_{t|t}^k = \hat{G'}_{t|t-1}^k + K_t^k \nu_t^k \tag{79}$$

$$M_{t|t}^k = \left(I - K_t^k \begin{bmatrix} f_t^k \\ 1 \end{bmatrix} \right) M_{t|t-1}^k \tag{80}$$

$$= \left(\frac{(f_t^k)^2}{\sigma_{R_t^k}^2} + \frac{1}{Q_t^k} + \frac{1}{M_{t|t-1}^k} \right)^{-1} \tag{81}$$

We need to relate the error of this estimate $M_{t|t-1}^k$ and $M_{t|t}^k$ with error $P_{t|t-1}^k$ and P_t^k in above section. With the fact that

$$M_{t|t}^k = Cov(\overline{G}_t^k|U_{[t]}^k, G_{[t]}) \tag{82}$$

$$= Cov(\overline{G}_t^k - G_t^{des}|U_{[t]}^k, G_{[t]}^{des}) \tag{83}$$

$$= Cov(\overline{G}_t^k - G_t^{des}|U_{[t]}^k, G_{[t]}^{des}, \hat{G}_t^k) \tag{84}$$

$$= Cov(\overline{G}_t^k - G_t^{des}|G_t^{des} - \hat{G}_t^k) \tag{85}$$

We can let $G_t^{des} - \overline{G}_t^k \perp \overline{G}_t^k - \hat{G}_t^k$ because it is existed when choosing some parameters. Let $\hat{G}_t^k = b_t^k U_t^k + B_t^k U_{[t-1]}^k$ and $\overline{G}_t^k = c_t^k \tilde{G}_t^k + C_t^k \tilde{G}_{[t-1]}^k$ with (68), the independent relationship is true when choosing $c_t^k = 1$, $b_t^k \Psi_t^k - C_t^k = b_t^k \Gamma_t^k + B_t^k = \mathbf{0}$ and $h_t^k b_t^k - c_t^k = 0$. Therefore

$$M_{t|t}^k = Cov(\overline{G}_t^k - G_t^{des}|G_t^{des} - \hat{G}_t^k) \tag{86}$$

$$= Cov(\overline{G}_t^k - G_t^{des})(1 - \frac{Cov(\overline{G}_t^k - G_t)}{Cov(G_t - \hat{G}_t^k)}) \tag{87}$$

$$= Q_t^k (1 - \frac{Q_t^k}{P_{t|t}^k}) \tag{88}$$

Similarly, $M_{t|t-1}^k$ can be related to $P_{t|t-1}^k$ in the same way as

$$Cov(\overline{G}_t^k|U_{[t-1]}^k, G_{[t]}) = Q_t^k (1 - \frac{Q_t^k}{P_{t|t-1}^k}) \tag{89}$$

Putting (66) and (67) into (71) to finish the proof of this theorem. ■

Combine the results in above would achieve Theorem 1 with the fact that function (22) is obtained directly from Theorem 3 and Theorem 4. The condition (23) is to guarantee the covariance in (89) greater than zero. The condition (24) is from (63) with the variance $\sigma_{V_t^k}^2$ is greater than zero. And the last condition (25) is due to the constraint of distortion.

B. Theoretical Analysis of Algorithm 1

Firstly, we turn the optimal problem (P2) into the following optimal problem with the same optimal $P_{t|t}^k$ to achieve rate-distortion function

$$(P3) \quad R_t'(P_{t|t}^{[k]}) = \inf_{P_{t|t}^{[K]}} \frac{1}{2} \log(\sum_{k=1}^K \frac{1}{P_{t|t}^k} - \frac{K-1}{\sigma_{G_t^{des}}^2}) \tag{90}$$

$$- \sum_{k=1}^K \frac{1}{2} \log(1 - \frac{Q_t^k}{P_{t|t}^k})$$

$$s.t. \quad P_{t|t}^k > Q_t^k \tag{91}$$

$$P_{t|t}^k \geq (\frac{1}{\sigma_{N_t^k}} + \frac{1}{P_{t|t-1}^k})^{-1} \tag{92}$$

$$\sum_{k=1}^K (\frac{1}{P_{t|t-1}^k} - \frac{1}{\sigma_{G_t^{des}}^2}) + \frac{1}{\sigma_{G_t^{des}}^2} \geq \frac{1}{D_t} \tag{93}$$

The function of $R_t'(P_{t|t}^{[k]})$ is the convex of the factor $P_{t|t}^{[k]}$. Then, we turn it into Lagrangian function

$$L(D_t) = \log \left[\sum_{j \neq k} (\frac{1}{P_{t|t}^j} - \frac{1}{G_t^{des}}) + \frac{1}{P_{t|t}^k} \right] \tag{94}$$

$$- \sum_{k=1}^K \log \left[1 - \frac{Q_t^k}{P_{t|t}^k} \right] \tag{95}$$

$$- \sum_{k=1}^K \lambda_t^k \left[P_{t|t}^k - \left(\frac{1}{\sigma_{N_t^k}^2} + \frac{1}{P_{t|t-1}^k} \right)^{-1} \right] \tag{96}$$

$$- \nu_t \left[\sum_{j \neq k} (\frac{1}{P_{t|t}^j} - \frac{1}{G_t^{des}}) + \frac{1}{P_{t|t}^k} - \frac{1}{D_t} \right] \tag{97}$$

$$s.t. \quad P_{t|t}^k > Q_t^k \tag{98}$$

Assume that $\Gamma_t^k = \sum_{j \neq k} (\frac{1}{P_{t|t}^j} - \frac{1}{G_t^{des}})$

The KKT condition is

$$\frac{\partial L(D_t)}{\partial P_{t|t}^k} = 0 \tag{99}$$

$$\lambda_t^k \left[P_{t|t}^k - \left(\frac{1}{\sigma_{N_t^k}^2} + \frac{1}{P_{t|t-1}^k} \right)^{-1} \right] = 0 \tag{100}$$

$$\nu_t \left[\Gamma_t^k + \frac{1}{P_{t|t}^k} - \frac{1}{D_t} \right] = 0 \tag{101}$$

$$P_{t|t}^k > Q_t^k \tag{102}$$

(a) If $P_{t|t}^k > \left(\frac{1}{\sigma^2_{N_t^k}} + \frac{1}{P_{t|t-1}^k}\right)^{-1}$, $\lambda_t^k = 0$. Then from (99), we can get that

$$\nu_t = \frac{1}{\Gamma_t^k + \frac{1}{P_{t|t}^k}} + \frac{Q_t^k}{1 - Q_t^k \frac{1}{P_{t|t}^k}} \tag{103}$$

$$\frac{1}{P_{t|t}^k} = \frac{1}{2}\left[\left(\frac{1}{Q_t^k} - \Gamma_t^k\right) + \sqrt{\left(\frac{1}{Q_t^k} + \Gamma_t^k\right)^2 - \frac{4\left(\frac{1}{Q_t^k} + \Gamma_t^k\right)}{\nu_t}} \right] \tag{104}$$

(b) If $P_{t|t}^k = \left(\frac{1}{\sigma^2_{N_t^k}} + \frac{1}{P_{t|t-1}^k}\right)^{-1}$, $\lambda_t^k = \frac{1}{(P_{t|t}^k)^2}(\nu - \frac{1}{\Gamma_t^k + \frac{1}{P_{t|t}^k}} - \frac{Q_t^k}{1 - Q_t^k \frac{1}{P_{t|t}^k}})$. Implying that $\nu \geq \frac{1}{\Gamma_t^k + \frac{1}{P_{t|t}^k}} - \frac{Q_t^k}{1 - Q_t^k \frac{1}{P_{t|t}^k}}$

Then, we come to the iterative water-filling algorithm

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] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[3] 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.

[4] 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.

[5] 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.

[6] 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, p. 17091720.

[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] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. on Signal Proce.*, vol. 68, pp. 2128–2142, 2020.

[9] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1508–1518.

[10] 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.

[11] 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.

[12] 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.

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

[14] ——, "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.

[15] Y. Yang, Z. Zhang, and Q. Yang, "Communication-efficient federated learning with binary neural networks," *IEEE J. Sele. Areas Commun.*, vol. 39, no. 12, pp. 3836–3850, 2021.

[16] 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. Syst.*, 2017, pp. 1709–1720.

[17] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Representations*, 2018.

[18] B. McMahan, H, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," in *Proc. Artif. Intell. Statist.(AISTATS)*, 2017, pp. 1–10.

[19] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. on Signal Proce.*, vol. 69, pp. 500–514, 2021.

[20] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, "Stochastic distributed learning with gradient quantization and variance reduction," 2019. [Online]. Available: arXiv:1904.05115

[21] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "SignSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 560–569.

[22] S. Horváth, C.-Y. Ho, L. Horváth, A. N. Sahu, M. Canini, and P. Richtarik, "Natural compression for distributed deep learning," preprint, 2019. [Online]. Available: arXiv:1905.10988

[23] P. Mayekar and H. Tyagi, "RATQ: A universal fixed-length quantizer for stochastic optimization," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 3130–3154, 2021.

[24] A. L. Perryman, D. Inoyama, J. S. Patel, S. Ekins, and J. S. Freundlich, "Pruned machine learning models to predict aqueous solubility," *ACS Omega*, vol. 5, no. 27, pp. 16 562–16 567, 2020.

[25] C. Hardy, E. Le Merrer, and B. Sericola, "Distributed deep learning on edge-devices: Feasibility via adaptive compression," in *IEEE Int. Symp. Netw. Comput. Appl.*, 2017, pp. 1–8.

[26] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2017, pp. 440–445.