NQFL: Nonuniform Quantization for Communication Efficient Federated Learning

Guojun Chen, *Student Member, IEEE,* Kaixuan Xie, Yuheng Tu, Tiecheng Song, *Member, IEEE,* Yinfei Xu, *Member, IEEE,* Jing Hu, *Member, IEEE* and Lun Xin

***Abstract*—Federated learning (FL), as a potential machine learning framework for privacy preservation, has gained significant attention. However, the considerable communication overhead associated with FL remains a prominent challenge. To mitigate this issue, a nonuniform quantization scheme based on Lloyd-Max algorithm is introduced in this letter. By employing this approach, less communication resources are consumed to achieve the same performance. Through performance analysis and numerical simulations, we verify the convergence and effectiveness of the proposed algorithm. It demonstrates the potential of our approach in reducing communication overhead while maintaining reliable performance in FL systems.**

***Index Terms*—Federated Learning, Communication Efficiency, Nonuniform Quantizer, Lloyd-Max Algorithm.**

I. Introduction

AMong the various machine learning (ML) methods, federated learning (FL) emerges as a prominent candidate for preserving privacy, initially introduced by the research team at Google [1]. In FL, local clients cooperate to implement learning. Specifically, they are allowed to jointly train a global ML model without sharing raw data among themselves or transmit raw data to a central server [2].

However, the constant updating of model parameters may engender substantial communication overload, which becomes a major bottleneck in specific scenarios [3]. In pursuit of higher communication efficiency, series of well-regarded strategies come to the fore. These include model updates size reduction and communication frequency reduction, which encompasses techniques such as quantization [4]–[8], sparsification , and periodic aggregation. Among these, quantization stands as a preeminent choice owing to its adaptability and robust performance [9].

The process involves applying lossy compression algorithms to quantize each entry to a finite-bit low-precision value. However, existing analyses have predominantly revolved around the use of uniform quantizers [4], [5], [10] across various scenarios and quantization levels. Additionally, renowned for generating a globally optimal quantizer when the probability density function (PDF) of a random variable is log-concave, the Lloyd-Max algorithm is recognized for scalar quantization [11], [12]. Although [6]–[8] provided compression scheme based on Lloyd-Max algorithm, [6] only normalizes the value of each parameters into $[0, 1]$. All of them require local clients and central server to generate and transmit the parameters of quantizers per iterative round, which incurring substantial computational and communication costs.

There is a scarcity of existing works that consider the statistical characteristics of the transmitted data to enhance compression. Recent research on gradient distribution analysis [13]–[15] demonstrates that the distribution of coordinates in each local gradients vectors tends to a Gaussian distribution. Prior works, such as [14], [16], have introduced rate-distortion theory to mitigate gradient redundancy between local clients. However, their compression methods come at the cost of computational complexity and local privacy.

To overcome these challenges, we propose a nonuniform quantization FL(NQFL) approach, offering greater flexibility and simplicity. The objective of our work is to propose an encoding-decoding system that reduces resource consumption and mitigates the impact of quantization errors, enabling the FL system effectively complete the learning process. The main contributions of this letter are summarized as:

- The proposed NQFL scheme utilizes a data normalization method to ensure that the transmitted data can use the same quantizer, thereby reducing the computational and communication overhead caused by frequently generating quantizers and transmitting quantizer parameters.
- The proposed NQFL utilizes the statistical characteristics of local gradients and compresses them in a nonuniform quantizer based on Lloyd-Max algorithm. Thereby, the quantized error is mitigated to improve the convergence speed of FL, decreasing the iterative rounds and communication costs.
- The convergence of the proposed NQFL scheme is analyzed theoretically, and the performance of NQFL is evaluated numerically.

The rest of this paper is structured as follows. Section II presents the system model of FL. In Section III, we design the nonuniform quantizer based on the data pre-processing method and Lloyd-Max algorithm. The performance analysis is established in Section IV. Section V contains the numerical simulation results, and Section VI contains the conclusion.

II. System Model

In this section, we provide a comprehensive overview of the framework of the communication efficient FL system. The

Guojun Chen, Tiecheng Song and Jing Hu are with the School of Information Science and Engineering, Southeast University, Nanjing, 210096, China, and also with the National Mobile Communication Research Laboratory, Nanjing 210096, China (e-mail: guojunchen@seu.edu.cn; songtc@seu.edu.cn; louy@seu.edu.cn). Kaixuan Xie and Lun Xin are with the China Mobile Research Institute, Beijing 100053, China (e-mail:xiekaixuan@chinamobile.com; xinlun@chinamobile.com). Yuheng Tu and Yinfei Xu are with the School of Information Science and Engineering, Southeast University, Nanjing, 210096, China (e-mail: 213213274@seu.edu.cn; yinfeixu@seu.edu.cn).

FL system comprises K local clients and a central server that cooperatively establish models in a distributed manner. Similar to the standard FL, The objective is to recover the optimal $m \times 1$ model vector $\boldsymbol{\omega}^o$ that satisfies

$$\boldsymbol{\omega}^o = \arg\min_{\boldsymbol{\omega}} \left\{ F(\boldsymbol{\omega}) \triangleq \frac{1}{K} \sum_{k=1}^{K} F_k(\boldsymbol{\omega}) \right\}. \tag{1}$$

where $\{\boldsymbol{x}_k^i, \boldsymbol{y}_k^i\}_{i=1}^{n_k}$ denote the set of n_k labeled data samples available at the kth local client, with $k \in \{1, 2, \ldots, K\}$. The local objective functions are defined as the empirical average over the training set

$$\begin{aligned} F_k(\boldsymbol{\omega}) &\triangleq F_k(\boldsymbol{\omega}; \{\boldsymbol{x}_k^i, \boldsymbol{y}_k^i\}_{i=1}^{n_k}) \\ &\triangleq \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(\boldsymbol{\omega}; \{\boldsymbol{x}_k^i, \boldsymbol{y}_k^i\}), \end{aligned} \tag{2}$$

where $\ell(\cdot)$ represents the sample-wise loss function that measures the prediction error of the model $\boldsymbol{\omega}$ on the training samples x^{n_k} with respect to the labels y^{n_k}. In this paper, for communication efficient FL, the process of each iteration is divided into three phases: local updated phase, model transmission phase and global aggregating phase. This paper focus on designing an efficient scheme for model transmission phase since another two phases are similar to the standard FL.

Local Updated Phase: After downloading the global model parameters $\boldsymbol{\omega}$ from the central server, each local client trains a specific neural network model by minimize the objective function $F_k(\boldsymbol{\omega})$. Due to the typically large number of model parameters $\boldsymbol{\omega}$ and the intrinsic complexity of most machine learning models, finding a closed-form solution to the optimization problem (1) is usually intractable [6]. Thus, local clients calculate the model gradients based on stochastic gradient descent (SGD) to solve this problem, expressed as

$$\boldsymbol{g}_k \triangleq \nabla F_k(\boldsymbol{\omega}). \tag{3}$$

where, $\boldsymbol{g}_k$ is denoted as the gradients on the kth local client. For distributed collaborative model training, the gradients need to be transmitted to the global server through model transmission phase.

Model Transmission Phase: In communication efficient FL, this process is crucial due to the precious and scarce nature of communication resources. It is imperative to compress the gradients into a finite-bit representation before transmission. Depending on the channel status, the updated gradients are need to be transmitted with a limited bit rate R, which is determined by the global server. The model transmission phase consists of two steps: encoding and decoding.

The encoding step is performed on the local clients. In this step, each coordinate of $\boldsymbol{g}_k \in \mathbb{R}^{n_k}$ is encoded into a digital codeword of R bits, represented as $u_k \in \{0, \cdots, 2^R - 1\} \triangleq \mathcal{U}_k$. Given the ith element of gradients $g_k^i \in \mathcal{R}$, where $i \in \{1, 2, \ldots, n_k\}$, the encoding function is expressed as:

$$e_k : \mathcal{R} \to \mathcal{U}_k. \tag{4}$$

Afterwards, each local client uploads the codewords $\{u_k^i\}_{i=1}^{n_k}$ to the central server under a certain communication rate R.

The decoding phase is processed on the central server. Upon receiving the codewords $\{u_k^i\}_{i=1}^{n_k}$ from all local clients, the estimated $\hat{\boldsymbol{g}}_k$ is reconstructed using decoding functions:

$$d_k : \mathcal{U}_k \to \mathcal{R}. \tag{5}$$

Global Aggregating Phase: Then the central server aggregates the global gradient by federated averaging (FA), via

$$\boldsymbol{g} = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{g}}_k, \tag{6}$$

and updates the current global model as

$$\boldsymbol{\omega}^{update} = \boldsymbol{\omega} - \eta \boldsymbol{g}, \tag{7}$$

where η is the learning rate. Finally, the updated global gradient $\boldsymbol{g}$ is broadcast to all the local clients. In addition, the central server jointly consider the channel status and loss function value to adjust transmission rate R of the next iteration.

III. Nonuniform Quantized FL Design

In this section, we propose the NQFL design in detail, mainly employed in the model transmission phase. The NQFL process consists three steps: data preprocessing, data quantization and data estimation. This section will explain the above three stages one by one. Then we will design NQFL to solve the learning problem expressed in (1) by minimizing the quantization distortion and decreasing the computation time.

A. Data Preprocessing

In order to better employ the compression algorithms, an effective data pre-processing method is crucial for maximizing their efficiency. A common practice is to normalize each gradients component g_k^i to the range $[0, 1]$ using the formula $\frac{|g_k^i|}{||\boldsymbol{g}_k^i||}$, as widely applied in [5], [6]. However, each the local client needs to design a specific quantizer and share the parameters of the quantizer with the central server per iteration. This can lead to unnecessary computational and communication overheads.

In the proposed NQFL system, we utilize the property of the gradients for some specific learning models. Inspirited by [13]–[15], the distribution of each coordinate g_k^i in local gradient vectors $\boldsymbol{g}_k$ tends to a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, where μ_k and σ_k are the mean and standard deviation of $\{g_k^i\}_{i=1}^{n_k}$, respectively. To enable all participants use a unified quantizer, the gradients are normalized to the standard normal distribution $\mathcal{N}(0, 1)$ by

$$\tilde{g}_k^i = \frac{g_k^i - \mu_k}{\sigma_k} \tag{8}$$

Thus, the local clients only need to upload the quantized value of gradients with mean and variance instead of the parameters of quantizers. In addition, the local clients have no need to update the quantizers if the required transmission rate is unchanged. This may significantly reduce computational overhead. For instance, the computational complexity for proposed normalized processing is $\mathcal{O}(n_k)$. Since the computational complexity of generating a Lloyd-Max quantizer is $\mathcal{O}(n \cdot 2^R \cdot$

T_{LM}), where T_{LM} denotes the number of the interative round for Lloyd-Max algorithm. The total computational complexity of the proposed NQFL is $T_{FL} \cdot K(\mathcal{O}(n)) + \mathcal{O}(n \cdot 2^R \cdot T_{LM})$, while the methods based on the standard Lloyd-Max algorithm cost $T_{FL} \cdot K(\mathcal{O}(n) + \mathcal{O}(n \cdot 2^R \cdot T_{LM}))$, where T_{FL} represent the number of iterative rounds for FL. Thus, the computational complexity of NQFL can decrease $(T_{FL} \cdot K - 1)\mathcal{O}(n \cdot 2^R \cdot T_{LM})$.

B. Data Quantization

With the aforementioned data pre-processing method, the PDF of gradient components is normalized to $\mathcal{N}(0, 1)$. Therefore, the nonuniform quantizer is designed based on the Lloyd-Max algorithm in this paper.

Firstly, we provide a brief overview of the Lloyd-Max algorithm, which shares similarities outlined in [17, Section III.A]. Generally, an M-level quantizer $Q(x)$ for an input value x consists of a pair set of quantization levels $\boldsymbol{l} = l_1, \cdots, l_M$ and quantization boundaries $\boldsymbol{\tau} = \tau_0, \cdots, \tau_M$. The quantization result $Q(x) = l_m$ if $x \in (\tau_{m-1}, \tau_m]$, $m \in 1, \cdots, M$. To minimize MSE, the Lloyd-Max algorithm determines the optimal quantization parameters based on the following iterative process:

$$l_m = \frac{\int_{\tau_{m-1}}^{\tau_m} x f_X(x) \mathrm{d}x}{\int_{\tau_{m-1}}^{\tau_m} f_X(x) \mathrm{d}x}, \tag{9}$$

$$\tau_m = \frac{(l_{m-1} + l_m)}{2}, \tag{10}$$

where $f_X(x)$ is the PDF of input x. The distortion is defined by the MSE as

$$D = \sum_{m=1}^{M} \int_{\tau_{m-1}} \tau_m (x - l_m)^2 f_X(x) \mathrm{d}x. \tag{11}$$

Therefore, with the given transmission rate R, the optimal quantizer with $M = 2^R$ levels is calculated by the aforementioned Lloyd-Max algorithm until the distortion convergence. We denote the optimal parameters of quantization levels and quantization bounds as $\boldsymbol{l}^* = \{l_1^*, \cdots, l_M^*\}$ and $\boldsymbol{\tau}^* = \{\tau_0^*, \cdots, \tau_M^*\}$, respectively. The nonuniform quantization function performs

$$Q_{NQ}(x, R) = l_m^* \text{ if } x \in (\tau_{m-1}^*, \tau_m^*] \tag{12}$$

After presenting the quantization scheme, we discuss the communication costs for one learning iteration round. It requires $R = \lceil log_2 M \rceil$ bits to represent the quantized index of each coordinate $\tilde{g}_k^i$. To mitigate the quantization bias, the mean and variance are transmitted with full precision using 32 bits, which can be transmitted in block with bit-limited. Therefore, denoting the total number of bits as C_s, the total communication costs of the FL process are

$$C_s = \sum_{t=1}^{T} \Big(\sum_{k=1}^{K} (n_k \lceil log_2 M \rceil) + 64 \Big) \tag{13}$$

where, T and K is the total iterative round and number of clients. n_k is the dimensions of the gradients on the kth client.

Remark 1: For each iterative round, the communication costs for each client are $n_k \lceil log_2 M \rceil) + 64$ bits. However, compared with other compression method, the communication costs required in [6] are $C_{SLM} = n_k \lceil log_2 M \rceil + n_k + 32 + 32(2M + 1)$, where $32(2M + 1)$ is used to transmit the quantizer with M levels and $M+1$ bounds. Thus, our proposed NQFL scheme decreases about $T \cdot K \cdot (n_k + 64M)$ bits, which is non-negligible since the number of labels is usually considerably larger (e.g., millions) [7].

C. Data Estimation

After receiving the codewords and pre-processing parameters, the global server reconstructs each coordinate g_k^i via

$$\hat{g}_k^i = Q_{NQ}^{-1}(u_k^i, R) \cdot \sigma_k + \mu_k, \tag{14}$$

where, $Q_{NQ}^{-1}(u_k^i, R)$ denotes the inverse function of quantization. Then, the central server aggregates the model parameters by (7) and computes the communication rate R_{t+1} for the next iteration. Finally, broadcasting the updated model parameters and the communication rate to finish one learning round.

D. NQFL Design

With the aforementioned nonuniform quantization scheme, we are now ready to introduce our NQFL algorithm in Algorithm 1.

Algorithm 1 Proposed NQFL algorithm

Input: Number of local clients K; Data samples $\{\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)}\}_{i=1}^{n_k}$; Preset loss threshold ξ.

1: **while** $F(\boldsymbol{\omega} > \xi)$ **do**
2: **for** $k = 1$ to K **do**
3: **Center server:** Broadcast the current model $\boldsymbol{\omega}$ and communication rate R to all local clients.
4: **Local clients:** Train the local model and calculate the gradient $\boldsymbol{g}_k = \{g_k^i\}_{i=1}^{n_g}$ by (1)-(3).
5: Pre-process the gradients by (8).
6: Quantize each coordinate to $Q_{NQ}(\tilde{g}_k^i, R)$ by (12).
7: Transmit the μ_k, σ_k and u_k^i to the global server.
8: **Center server:** Recover gradients $\{\hat{\boldsymbol{g}}_k\}_{k=1}^{K}$ and compute the aggregated model according to (6), (7).
9: **end for**
10: **end while**

The main idea of NQFL operates at the local clients, as depicted in lines 3-6 of Algorithm 1. To reduce computational overhead by using standardized quantizers, each local client normalizes the coordinate of gradients as $\tilde{g}_k^i \sim \mathcal{N}(0, 1)$ during data pre-processing. Consequently, local clients quantize the gradients using the proposed nonuniform quantizer defined in (12). They only need to transmit two additional parameters and the index of each quantized gradient without the parameters of quantizer. Finally, the central server decodes the gradients and aggregates them using FA to finish one iterative round.

IV. Performance Analysis

In this section, we proceed to analyze the performance of NQFL by examining its distortion characteristics and studying its convergence properties. The analysis is conducted based on

the following assumptions, which are commonly employed in convergence studies of FL [4], [5], [18]:

ASS1: The expected squared norm of stochastic gradients $\boldsymbol{g}_k$ in (3) is uniformly bounded by some $G_k^2 > 0$, i.e. $\mathbb{E}||\boldsymbol{g}_k||^2 \leq G_k^2$.

ASS2: The second moment of stochastic gradient for all function $F_k(\cdot)$ is bounded: $\mathbb{E}[(\boldsymbol{g}_k - \mathbb{E}[\boldsymbol{g}_k])^2] \leq \Sigma_k$.

ASS3: The local objective functions $F_k(\cdot)$ are all L-smooth: for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{R}_m$, it holds that: $F_k(\boldsymbol{v}_1) - F_k(\boldsymbol{v}_2) \leq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F_k(\boldsymbol{v}_2) + \frac{L}{2}||\boldsymbol{v}_1 - \boldsymbol{v}_2||_2^2$.

ASS4: The local objective functions $F_k(\cdot)$ are all C-strongly convex: for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{R}_m$, it holds that: $F_k(\boldsymbol{v}_1) - F_k(\boldsymbol{v}_2) \geq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F_k(\boldsymbol{v}_2) + \frac{C}{2}||\boldsymbol{v}_1 - \boldsymbol{v}_2||_2^2$.

The convergence properties of NQFL with federated averaging is characterized in the following theorem:

Theorem 1: Let Assumptions ASS1-ASS4 hold and L, C, G_k and Σ_k be defined therein. Choose $\gamma = \max\{8\frac{L}{C}, E\} - 1$ and the learning rate $\eta_t = \frac{2}{C(t+\gamma)}$. By denoting ω^{IN} as the initial model parameters, it satisfies

$$\mathbb{E}\{F(\boldsymbol{\omega}^t) - F(\boldsymbol{\omega}^o)\} \leq \frac{L}{2(t+\gamma)} \max\left\{\frac{4B}{C^2}, (\gamma+1)\mathbb{E}\{||\boldsymbol{\omega}^{IN} - \boldsymbol{\omega}^o||^2\}\right\},$$

where

$$B = \frac{\sqrt{3}\pi}{2K}\sum_{k=1}^{K} 2^{-2R}\sigma_k^2 + 6L\phi + \frac{8(E-1)}{K}\sum_{k=1}^{K} G_k^2, \quad (15)$$

E is the retraining times of $\boldsymbol{\omega}_k^t$ by local clients and $\phi = F(\boldsymbol{\omega}^o) - \frac{1}{K}\sum_{k=1}^{K} \min_{\boldsymbol{\omega}} F_k(\boldsymbol{\omega})$ is the heterogeneity gap.

Remark 2: These theorem implies that our proposed NQFL algorithm converges at a rate of $\mathcal{O}(\frac{1}{t})$. Specifically, the difference between the objective of the model learned in a federated manner and the optimal objective decays to zero at least as quickly as $\frac{1}{t}$, which is the same order of convergence as FL [5], [18].

Structural Proof: First of all, we expand the averaged noisy stochastic gradients in (6) and define the desired averaged gradients as

$$\boldsymbol{g}^t = \frac{1}{K}\sum_{k=1}^{K}(\nabla F_k(\boldsymbol{\omega}^t) - \frac{1}{\eta_k^t}\epsilon_k), \quad (16)$$

$$\boldsymbol{g}_{des}^t := \frac{1}{K}\sum_{k=1}^{K}\nabla F_k(\boldsymbol{\omega}^t). \quad (17)$$

Subsequently, we refer to the resulting model used in [18, Lemma 1]. With the assumptions ASS3 and ASS4, if $\eta_t \leq \frac{1}{4L}$, the expected distance between the ML model parameters $\boldsymbol{\omega}^{t+1}$ in (7) and the optimal parameters $\boldsymbol{\omega}^o$ in (1) is bounded by

$$\mathbb{E}\{||\boldsymbol{\omega}^{t+1} - \boldsymbol{\omega}^o||^2\} \leq (1 - \eta_t C)\mathbb{E}\{||\boldsymbol{\omega}^t - \boldsymbol{\omega}^o||^2\} + 6L\eta_t^2\phi + 2\underbrace{\mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K}||\boldsymbol{\omega}^t - \boldsymbol{\omega}_k^t||^2\right\}}_{(a)} + \eta_t^2\underbrace{\mathbb{E}\{||\boldsymbol{g}^t - \boldsymbol{g}_{des}^t||^2\}}_{(b)}, \quad (18)$$

where $\boldsymbol{\omega}_k^t = \boldsymbol{\omega}_k^{t-1} - \eta_{t-1}\boldsymbol{g}_k^t$ is updated ML model parameters on the kth UE To further bound Eq.(18), we then derive the bounding of gradients and the divergence of $\boldsymbol{\omega}_k^t$. Then, we bound the items (a) and (b) of Eq.(18) respectively.

For the divergence of $\boldsymbol{\omega}_t^k$, shown as (a) in Eq.(18), it is bound with ASS1 and the fact that η_n is non-increasing, via

$$\mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K}||\boldsymbol{\omega}_k^t - \boldsymbol{\omega}^t||\right\} \leq 4(E-1)^2\eta_t^2\frac{1}{K}\sum_{k=1}^{K}G_k^2, \quad (19)$$

The derivation is similar to [5, Lemma C.2] and [18, Lemma 3], we omit it here due to the limitation of page space.

Then, for the item (b), the approximate quantization error variance in a M-level Lloyd-Max quantizer [19] is

$$\mathbb{E}[(x - \hat{x})^2] \approx \frac{1}{12M^2}\left[\int_x \sqrt[3]{f_X(x)}dx\right]^3. \quad (20)$$

Here, since each coordinate $g_k^i \sim \mathcal{N}(\mu_k, \sigma_k^2)$ and number of quantization levels is 2^R, with Panter-Dite approximation [19], the quantized distortion of g_k^i is

$$\mathbb{E}[(g_k^i - \hat{g}_k^i)^2] \approx \frac{\sqrt{3}\pi}{2}2^{-2R}\sigma_k^2. \quad (21)$$

With the assumption ASS2, the quantization error is bounded

$$\mathbb{E}[||\boldsymbol{g}^t - \boldsymbol{g}_{des}^t||^2] \approx \frac{1}{K}\sum_{k=1}^{K}\frac{\sqrt{3}\pi}{2}2^{-2R}\sigma_k^2. \leq \frac{\sqrt{3}\pi}{2K}\sum_{k=1}^{K}2^{-2R}\sigma_k^2. \quad (22)$$

Therefore, by defining $\Delta_t = \mathbb{E}||\boldsymbol{\omega}^t - \boldsymbol{\omega}^o||^2$, Eq.(18) is bounded by combining Eq.(19) and Eq.(22) via

$$\Delta_{t+1} \leq (1 - \eta_t C)\Delta_t + \eta_t^2 B. \quad (23)$$

Subsequently, for non-increasing $\eta_t = \frac{\beta}{t+\gamma}$ with some $\beta > \frac{1}{C}$ and $\gamma > 0$, when denoting $\nu = \max\{\frac{\beta^2 B}{\beta C - 1}, (\gamma+1)\delta_0\}$, the value of Δ_t is bounded by

$$\Delta_t \leq \frac{\nu}{t+\gamma}. \quad (24)$$

This derivation can be proved by mathematical induction, that it holds for $t = 1$, and it also holds for $t = t + 1$ according simple derivation from Eq.(23).

Finally, according to ASS3, when setting $\beta = \frac{2}{C}$ and combining above parameters, this theorem is proved.

V. Numerical Evaluations

In this section, we conduct a numerical evaluation of NQFL for a communication efficient FL system. The performance is evaluated in terms of comparing the costs of communication resources with a certain test accuracy.

We first introduce the setup of the numerical evaluations. The FL system consists of $K = 10$ local clients.We utilize the dataset of MNIST, FeMNIST and Cifar10. These datasets are partitioned into a training set of $60,000$, $805,263$ and $50,000$ samples, respectively. The deployment of models includes LR and CNN models, which contain $28 \times 28 = 784$ and $1,664,650$ model parameters, respectively. To evaluate the performance of NQFL, we set the initialized learning rate to 0.03. The number of local iterations on clients is 5 for LR and 20 for CNN.

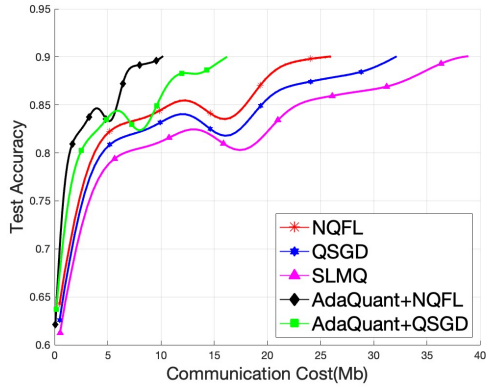

Fig. 1. Test accuracy versus communication costs on different algorithms

Both fixed and adaptive communication rate are considered in this section. We set the communication rate $R = 6$ bits for the condition of fixed quantization and initial $R = 1$ for the condition of adaptive quantization. The adaptive quantization model is based on [20], letting the number of quantized levels dynamically change according to the loss function. We compare the performance of NQFL against traditional QSGD [4] and up-to-date quantization algorithm based in Lloyd-Max quantizer, denoted as SLMQ, [6].

TABLE I
COMPARISON OF COMMUNICATION COSTS FOR DIFFERENT ALGORITHM WITH DIFFERENT DATASETS AND LEARNING MODELS WHEN OBTAINING 90% TEST ACCURACY

Dataset	Model	Algorithm	Communication Costs
MNIST	LR	NQFL	26.78Mb
		SLMQ	37.91Mb
		QSGD	32.41Mb
	CNN	AdaQ + NQFL	2.94Gb
		AdaQ + QSGD	3.41Gb
FeMNIST	CNN	NQFL	7.02Gb
		SLMQ	12.37Gb
		QSGD	8.32Gb
CIFAR10	CNN	AdaQ + NQFL	3.89Gb
		AdaQ + QSGD	4.86Gb

Fig. 1 illustrates a comparison of test accuracy versus the communication costs by NQFL, QSGD, and SLMQ, using the MNIST wiht LR model. The results clearly demonstrate the efficacy of NQFL algorithm. Regardless of a fixed or adaptive R, the proposed NQFL can achieve the accuracy (90%) with the least communication costs, reducing approximate 21.2% and 37.5% communication costs, respectively. Next, TABLE. I shows the consumption of communication resources for different algorithms under different datasets and learning models. When setting the required test accuracy be 90%, the proposed NQFL always maintains fewest communication overhead, saving 17.3%, 26.9%, 18.5% and 19.9% communication resources respectively. Consequently, the proposed NQFL algorithm outperforms both the uniform and non-uniform quantization algorithms in terms of both fixed and adaptive quantization scenario.

VI. CONCLUSION

In this paper, we present a novel nonuniform quantization scheme called NQFL, to effectively reduce communication overhead FL systems. By leveraging a novel data pre-processing method and the Lloyd-Max algorithm, NQFL enables compression of ML model to achieve higher communication efficiency. It is able to accelerate model convergence without adding much additional overhead. Furthermore, The convergence of NQFL is derived during the performance analysis. Finally, we evaluate the efficiency of NQFL using numerical experiments, demonstrating its superior performance compared to other quantization schemes.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 2017.

[2] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[3] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 492–523, 2021.

[4] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proceedings of the Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1709–1720.

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

[6] W. Liu, L. Chen, Y. Chen, and W. Wang, "Communication-efficient design for quantized decentralized federated learning," 2023. [Online]. Available: https://arXiv.org/abs/2303.08423

[7] S. Vargaftik, R. B. Basat, A. Portnoy, G. Mendelson, Y. B. Itzhak, and M. Mitzenmacher, "EDEN: Communication-efficient and robust distributed mean estimation for federated learning," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 21 984–22 014.

[8] F. Sheng, L. Xu-Jian, and Z. Li-Wei, "A lloyd-max-based non-uniform quantization scheme for distributed video coding," in *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, vol. 1, 2007, pp. 848–853.

[9] 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 Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1342–1397, 2021.

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

[11] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[12] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, 1960.

[13] N. Zhang, M. Tao, J. Wang, and F. Xu, "Fundamental limits of communication efficiency for model aggregation in distributed learning: A rate-distortion approach," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 173–186, 2023.

[14] H. Yang, T. Ding, and X. Yuan, "Federated learning with lossy distributed source coding: Analysis and optimization," 2022. [Online]. Available: https://arxiv.org/abs/2204.10985

[15] Y. Liu, S. Rini, S. Salehkalaibar, and J. Chen, "M22: A communication-efficient algorithm for federated learning inspired by rate-distortion," 2023. [Online]. Available: https://arxiv.org/abs/2301.092697

[16] G. Chen, L. Yu, W. Luo, Y. Xu, and T. Song, "Rate-distortion optimization for adaptive gradient quantization in federated learning," in *Processing of the IEEE Wireless Communications and Networking Conference (WCNC)*, Glasgow, United Kingdom, 2023.

[17] N. Nguyen-Thanh and I. Koo, "Log-likelihood ratio optimal quantizer for cooperative spectrum sensing in cognitive radio," *IEEE Communications Letters*, vol. 15, no. 3, pp. 317–319, 2011.

[18] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-IID data," in *Processing of the International Conference On Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[19] P. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.

[20] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3110–3114.