QMGeo: Differentially Private Federated Learning via Stochastic Quantization with Mixed Truncated Geometric Distribution

Zixi Wang, M. Cenk Gursoy
Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY
{zwang227, mcgursoy}@syr.edu

***Abstract*—Federated learning (FL) is a framework which allows multiple users to jointly train a global machine learning (ML) model by transmitting only model updates under the coordination of a parameter server, while being able to keep their datasets local. One key motivation of such distributed frameworks is to provide privacy guarantees to the users. However, preserving the users' datasets locally is shown to be not sufficient for privacy. Several differential privacy (DP) mechanisms have been proposed to provide provable privacy guarantees by introducing randomness into the framework, and majority of these mechanisms rely on injecting additive noise. FL frameworks also face the challenge of communication efficiency, especially as machine learning models grow in complexity and size. Quantization is a commonly utilized method, reducing the communication cost by transmitting compressed representation of the underlying information. Although there have been several studies on DP and quantization in FL, the potential contribution of the quantization method alone in providing privacy guarantees has not been extensively analyzed yet. We in this paper present a novel stochastic quantization method, utilizing a mixed geometric distribution to introduce the randomness needed to provide DP, without any additive noise. We provide convergence analysis for our framework and empirically study its performance.**

***Index Terms*—Federated learning, differential privacy, quantization.**

I. Introduction

The growing prevalence of machine learning (ML) applications underscores the critical need for a vast and diverse corpus of training data. Conventional centralized ML models, while effective, necessitate the central storage and processing of data, which raises significant privacy concerns and presents formidable challenges when dealing with devices and sensors equipped with limited data transmission capabilities and resources. In response to the above-mentioned challenges, federated learning (FL) has been intensively studied recently, as an approach where machine learning (ML) models of different users are jointly trained without requiring the sharing of the raw local data [1] [2]. Specifically, FL supports a distributed ML framework in which each user trains its own local ML model under the coordination of a central server, that combines the local models by requiring the users to transmit only their model updates. With the users' datasets remaining local, FL avoids the heavy communication cost and privacy leakage of transmitting entire datasets.

As mentioned above, one of the key motivations of FL structure is to reduce the communication cost. However, as modern ML applications start involving larger models, even transmitting only the model update at high precision can lay a heavy burden on communication resources, especially for edge devices whose transmission capabilities are often limited. This has led to quantization being widely used to overcome this challenge in FL. In particular, the updates from the users are first quantized into an efficient representation before the transmission [3] [4].

In addition to being a more communication-efficient scheme, another key motivation behind FL is that it alleviates privacy concerns to a degree by requiring only local processing of user data. However, while FL facilitates the users' data to remain decentralized at the users' local devices, this is not necessarily sufficient to ensure complete privacy. Recent studies have shown that by analyzing the model updates that get transmitted through the air [5] or just the final model itself [6] [7], privacy leakage can occur. Differential privacy (DP) [8] is almost a de-facto mechanism used for privacy guarantees in ML applications. DP provides privacy guarantees in FL systems by introducing random perturbation to the transmitted model update, thus introducing randomness and perturbation to conceal the sensitive information about user data. DP mechanism has already been implemented in many real world applications [9] [10] [11].

We note that quantization methods lead to a certain amount of distortion in the information that the transmitted model updates were able to encode. Since DP methods ensure privacy by introducing random perturbation, the next question follows naturally: If a quantization method is stochastic, would the quantization process itself be able to provide provable DP guarantee? Our work is primarily motivated by this question.

A. Related Work

There exist studies that jointly consider reducing communication cost by quantization and providing DP guarantees. For instance, the authors in [12] propose applying the binomial mechanism with quantization to achieve both communication efficiency and differential privacy, with an improved analysis showing better utility for the binomial mechanism. [13] efficiently discretizes and flattens the model updates from the users and adds discrete Gaussian noise. However, these mechanisms still rely heavily on additional noise injection to achieve differential privacy.

The authors in [14] introduce a novel mechanism where the model update information is first encoded into a parameter of a binomial distribution, and the mechanism generates samples from the distribution, without the need of additive noise. [15] also considers utilizing ternary quantization for privacy. However, their analysis applies to ternary quantization only and the DP guarantee they achieve is determined by the l-infinity norm of the model update.

B. Contributions

In this paper, we propose a novel randomized quantization method that provides DP guarantees for multiple configurations of the quantization method without relying on any additional noise injection. The main contributions of this paper are summarized as follows:

- We propose a novel stochastic scalar quantization method QMGeo that utilizes the geometric distribution and no other additional noise to achieve ϵ-DP, and we provide a Rényi DP analysis as well.
- We provide privacy analysis for the QMGeo method, when applied to scalar and vector models.
- We provide the optimality gap analysis to show how the FL framework converges.

II. Federated Learning Protocol

We first discuss an FL system with N clients and a parameter server (PS). The goal of such an FL system is to minimize a loss function $F(\boldsymbol{w})$ with regard to the model parameter $\boldsymbol{w} \in \mathbb{R}^d$. We denote the dataset that user m possesses as $\mathcal{B}_n$, and define B as the total number of samples in the system, i.e., $B = \sum_{n=1}^{N} |\mathcal{B}_n|$. We denote the user loss function as $F_n = \frac{1}{|\mathcal{B}_n|} \sum_{\boldsymbol{u} \in \mathcal{B}_n} f(\boldsymbol{w}, \boldsymbol{u})$, where $\boldsymbol{u}$ corresponds to one particular sample from the dataset $\mathcal{B}_n$. Thus, the total loss function is given as follows:

$$F(\boldsymbol{w}) = \sum_{n=1}^{N} \frac{|\mathcal{B}_n|}{B} F_n(\boldsymbol{w}). \tag{1}$$

An iterative approach is considered to minimize the above loss function. The PS first broadcasts the global model parameter w_t to all the users. The users apply a uniform sampling to acquire the mini batch used for training with a sampling rate $\kappa = \frac{b(n)}{|\mathcal{B}_n|}$, while $\boldsymbol{b}_t(n)$ is the set of samples with size b_n drawn from the dataset $\mathcal{B}_n$ in iteration t.

The user n acquires the gradient as

$$\boldsymbol{g}_t(n) = \nabla F_n(\boldsymbol{w}_t, \boldsymbol{b}_t(n)). \tag{2}$$

In addition, an element-wise clipping is applied to each element of the gradient $\boldsymbol{g}_t(n)$, where each element is clipped to be within the range of $[-W_{\max}, W_{\max}]$:

$$\bar{\boldsymbol{g}}_t(n) = clip(\boldsymbol{g}_t(n)). \tag{3}$$

The clipped gradient $\bar{\boldsymbol{g}}_t(n)$ is then quantized by a quantizer $Q(\cdot)$. The quantized vector $\Delta \boldsymbol{w}_t(n) = Q(\bar{\boldsymbol{g}}_n(t))$ is uploaded to the PS, which then aggregates all model updates from n users and update accordingly with a learning rate η:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \sum_{n=1}^{N} \Delta \boldsymbol{w}_n(t). \tag{4}$$

The updated global model parameter $\boldsymbol{w}_{t+1}$ is then broadcasted to all the users before the next iteration.

III. QMGeo: Quantization Method with Mixed Geometric Distribution

In this section we propose a stochastic scalar quantization method using a truncated mixed geometric distribution for the stochasticity. The quantization method considered is a scalar quantizer, quantizing the input scalar values, required to be limited within range $\{w | w \in [-W_{\max}, W_{\max}]\}$, into R quantization levels. The quantizer is parameterized by the number of quantization levels R and the parameter p for the geometric distributions used to generate the mixed geometric distribution. We define the quantizer as $Q_{\text{MGeo}}(\cdot) : \{w | w \in [-W_{\max}, W_{\max}]\} \rightarrow \{Bin(0), Bin(1), \ldots, Bin(R-1)\}$.

For each entry in the vector to be quantized, there would be a clipping threshold $W_{\max}$, limiting the value of the corresponding entry w to be within range $[-W_{\max}, W_{\max}]$. Then, for each integer $r \in [0, R)$, let a bin $Bin(r)$ represent the corresponding value for the r-th quantization level, such that the bins evenly span the range $[-W_{\max}, W_{\max}]$, where

$$Bin(r) = -W_{\max} + \frac{2rW_{\max}}{R-1}. \tag{5}$$

$Q_{\text{MGeo}}(w)$ takes a scalar w as input and outputs a value from the set of quantization levels $\{Bin(0), Bin(1), \ldots, Bin(r), \ldots, Bin(R-1)\}$ as the quantized output. The output value, i.e., the chosen quantization level $Bin(r)$, is sampled from a mixed truncated geometric distribution determined by the value of the input scalar w.

In the following part of this section, we first introduce the truncated geometric distribution, and then discuss how the truncated mixed geometric distribution is determined.

A. Truncated Geometric Distribution

The truncated geometric distribution is essentially a geometric distribution but supported only between the range of two truncation points $\{x | x \in \mathbb{Z}, x \in [a, b-1]\}$, where a and b are the left and right truncation points. In this paper, we consider only one right truncation, and thus the distribution is supported on $\{x | x \in \mathbb{Z}, x \in [1, b-1]\}$. As in [16], the probability mass function of a truncated geometric distribution $TGeo$ with a success probability p, a failure probability of $q = 1-p$ and a truncation point b is

$$Pr\{X_{TGeo} = k\} = p(1-p)^{k-1} \cdot \frac{1}{1-(1-p)^{b-1}}. \tag{6}$$

The mean and variance for the truncated geometric distribution $TGeo$ are

$$E(X_{TGeo}) = \frac{1 - bq^{b-1} + (b-1)q^b}{p(1-q^{b-1})}, \tag{7}$$

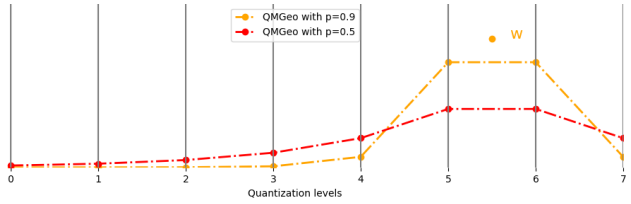

Fig. 1. How $\mathcal{Q}_{\text{MGeo}}(\cdot)$ quantizes a scalar input value w is demonstrated in this figure. The dotted line shows the probability of the input value w being quantized to the corresponding quantization level. As shown in the figure, each quantization level is assigned with a non-zero probability. The larger p is, the more skewed the distribution becomes.

$$\sigma^2_{TGeo} = \frac{(1+q^{2b})q - q^b(1+q^2)b^2 + q^{b+1}(b^2-1)}{(1-q)^2(1-q^b)^2}. \quad (8)$$

B. Mixed Truncated Geometric Distribution

We define a w centered mixed truncated geometric distribution as follows:

$$X_{\text{MTGeo}} = \begin{cases} Bin\left(r-(X_{TGeo1}-1)\right) & \text{w.p.} \quad p_{\text{mix}} \\ Bin\left(r+X_{TGeo2}\right) & \text{otherwise} \end{cases} \quad (9)$$

where $p_{\text{mix}} = \frac{Bin(r+1)-w}{Bin(r+1)-Bin(r)}$, and X_{TGeo1} and X_{TGeo2} are random variables sampled from two different truncated geometric distributions, $TGeo1(p)$ and $TGeo2(p)$. The mixture probability is determined by the distance from w to $Bin(r)$ and $Bin(r+1)$ respectively, where $Bin(r)$ and $Bin(r+1)$ are the neighboring quantization levels to w.

Since both the input and the output of the quantization scheme has a range of $[-W_{\text{max}}, W_{\text{max}}]$, we are forced to use truncated geometric distributions to sample the decided quantization level. Thus, $X \sim TGeo1(p)$ is supported only on $\{1,2,\ldots,r+1\}$ and $X \sim TGeo2(p)$ is supported only on $\{1,2,\ldots,R-r-1\}$. We obtain the truncated geometric distributions by normalizing the probability mass on the support back to 1, where the geometric distribution itself is parameterized by the success probability p:

$$Pr\{X_{TGeo1}=k\} = p(1-p)^{k-1}\cdot\frac{1}{1-(1-p)^{r+1}}, \quad (10)$$

$$Pr\{X_{TGeo2}=k\} = p(1-p)^{k-1}\cdot\frac{1}{1-(1-p)^{R-r-1}}. \quad (11)$$

The mixed geometric distribution assigns non-zero probabilities to every quantization level, providing the randomness needed to achieve differential privacy. We control the shape of the distribution with parameter p. In particular, the larger p is, the more skewed the distribution becomes, focusing the probability mass to neighboring quantization levels, thus less variance introduced by the $Q_{\text{MGeo}}(\cdot)$ mechanism, as demonstrated in Fig 1.

IV. Differential Privacy Analysis for QMGeo

We first give the following definitions regarding DP [8].

ϵ - differential privacy: For any two adjacent datasets $D, D^{'} \in \mathcal{D}$, and any output set of $\mathcal{S} \subset \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$, and randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ that satisfies (ϵ, δ) - differential privacy, the following property must hold:

$$Pr[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(D^{'}) \in \mathcal{S}]. \quad (12)$$

We note that more commonly, DP mechanism use the concept of (ϵ, δ)-DP, where δ is essentially a relaxation term that allows some arbitrarily small probability for the DP mechanism to fail.

(α, ϵ)-Rényi Differential Privacy (RDP): For any two adjacent datasets $D, D^{'} \in \mathcal{D}$, and any output set of $\mathcal{S} \subset \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$, and randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ that satisfies (α, ϵ)-RDP, the following must hold:

$$D_\alpha\left(P_{\mathcal{M}(D)} || P_{\mathcal{M}(D')}\right) \leq \epsilon, \quad (13)$$

where $D_\alpha\left(P_{\mathcal{M}(D)} || P_{\mathcal{M}(D')}\right)$ is the Rényi divergence, given by

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right]. \quad (14)$$

We first establish the guarantee that a scalar stochastic k-level quantization provides as a comparison. We next establish the per-element differential privacy the scalar QMGeo method provides in Section IV-B. We then extend the analysis to the differential privacy guarantee when applied to a model update vector.

A. Differential privacy of the stochastic k-level quantization

The stochastic k-level quantization is a quantization method that randomly assigns the value w to neighboring quantization levels, where the probability is determined by the distance from w to its neighboring quantization levels. The stochastic k-level quantization, similar to the $Q_{\text{MGeo}}(\cdot)$, requires the input scalar values to be limited within some range $\{w | w \in [-W_{\text{max}}, W_{\text{max}}]\}$ and the bins evenly span the range:

$$Bin(r) = -W_{\text{max}} + \frac{2rW_{\text{max}}}{R-1}. \quad (15)$$

The quantization is done as follows:

$$Q(w) = \begin{cases} Bin\left(r+1\right) & \text{w.p.} \quad \frac{w-Bin(r)}{Bin(r+1)-Bin(r)} \\ Bin\left(r\right) & \text{otherwise.} \end{cases} \quad (16)$$

For number of quantization levels larger than 3, the stochastic k-level quantization faces a difficulty providing DP, as it could only quantize values to their neighboring quantization levels. When analysing the DP performance, it is obvious that if g and $g^{'}$ do not lie in neighboring intervals, there will always be a choice of v that achieves the worst case rendering one of the probability terms to 0 and nonzero for the other. For instance, when v is a neighboring quantization level $v = Bin(r)$ to g but not for $g^{'}$, g will be quantized to v with a non-zero probability p_{mix}, while $g^{'}$ will never be quantized to v. In this case, ϵ cannot be bounded. An example is provided in Fig 2, where the intuition is that, since the quantization method quantizes to only neighboring

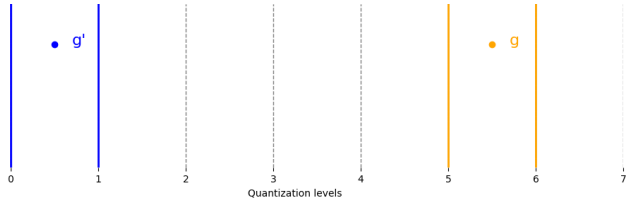

Fig. 2. This figure illustrates an example of the stochastic k-level quantization failing to provide DP. The solid orange and blue lines are the possible output quantization levels for g and $g^{'}$ respectively. It is obvious that, by observing the output of the quantization, any adversary could distinguish betwen g and $g^{'}$. Any grey dotted lines are quantization levels that lies neither in the range of $Q(g)$ or $Q(g^{'})$, thus any of them being chosen as v would render the ϵ unbounded.

quantization levels, there is not enough randomness to achieve DP for a number of quantization levels larger than 3:

$$e^{\epsilon} Pr[\mathcal{Q}(g^{'}) = v] \geq Pr[\mathcal{Q}(g) = v] \tag{17}$$

$$e^{\epsilon} \times 0 \geq p_{mix}. \tag{18}$$

However, for ternary quantization, such method could provide DP guarantees to some extent [15].

B. Differential privacy of the scalar QMGeo method

1) ϵ-DP: We first consider the ϵ-DP analysis. In our case, the considered randomized mechanism is essentially the quantization method that determines the vectors sent from the users' devices. Thus, we consider the following inequality:

$$Pr[\mathcal{Q}_{\text{MGeo}}(g) = v] \leq e^{\epsilon} Pr[\mathcal{Q}_{\text{MGeo}}(g^{'}) = v]. \tag{19}$$

(19) should hold for any v, where g and $g^{'}$ are entries of gradients acquired from arbitrary neighboring datasets of a particular user. We note that the worst case scenario occurs when v is either $Bin(0)$ or $Bin(R-1)$, g and $g^{'}$ are equal to $Bin(0)$ and $Bin(R-1)$ respectively. Without loss of generality, we assume $v = Bin(0)$, $g = Bin(0)$ and $g^{'} = Bin(R-1)$. Substituting the values in (19) with the probability mass of the mixed truncated geometric distribution, we obtain

$$e^{\epsilon} \geq \frac{Pr[Q_{\text{MGeo}}(g) = v]}{Pr[Q_{\text{MGeo}}(g^{'}) = v]} \tag{20}$$

$$e^{\epsilon} \geq \frac{\frac{1}{2}}{\frac{1}{2}p(1-p)^{R-2}\frac{1}{1-(1-p)^{R-1}}} \tag{21}$$

$$\begin{aligned}\epsilon &\geq -\ln p(1-p)^{R-2} + \ln\left(1-(1-p)^{R-1}\right) \\ &= -(\ln p + (R-2)\ln(1-p)) + \ln\left(1-(1-p)^{R-1}\right).\end{aligned}$$

We demonstrate the trade-off between the achieved DP level ϵ and the parameters of the $QMGeo(\cdot)$ i.e., the number of quantization levels r and the success probability p of the mixed truncated geometric distribution. Fig 3 demonstrates how the achieved DP level ϵ changes as the parameter p of $Q_{\text{QMGeo}}(\cdot)$ changes for fixed numbers of quantization levels. ϵ increases rapidly as p approaches 1, where at $p = 1$, the $Q_{\text{QMGeo}}(\cdot)$ reduces to a conventional stochastic r-level quantization which could not achieve DP except for when quantizing to 3 levels or less. As shown in (20), ϵ increases linearly with the number of quantization levels R.

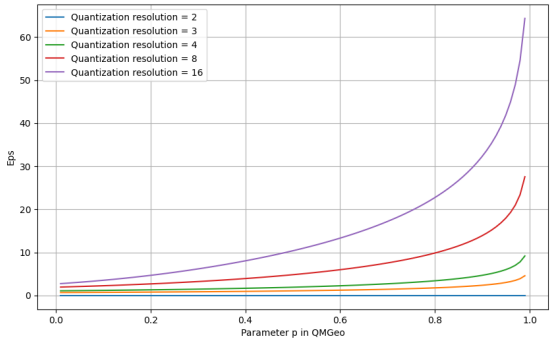

Fig. 3. ϵ-DP as a function of the parameter p of $Q_{\text{QMGeo}}(\cdot)$, the success rate of the mixed truncated geometric distribution. Each curve displayed in this figure corresponds to a different number of quantization levels.

2) (α, ϵ)-RDP: With the quasi-convex property of the Rényi divergence, we can show that the worst case scenario is attained by the extremal points where two input elements each sit at $Bin(0)$ and $Bin(R-1)$, where the two discrete distributions are essentially a standard truncated geometric distribution X and a slightly altered one, $X^{'} = -X + R + 1$, sharing the same parameter of success rate p and supported on the same set of quantization levels. We then substitute parameters R, p from the $\mathcal{Q}_{\text{MGeo}(\cdot)}$ into (13) and obtain

$$D_{\alpha}\left(P_{\mathcal{M}(D)} || P_{\mathcal{M}(D')}\right) \leq D_{\alpha}\left(P_X || P_{X'}\right) \tag{22}$$

$$\begin{aligned}&= \frac{1}{\alpha-1}\log\left\{\frac{1}{2}\frac{(1-q^{R-1})^{\alpha-1}}{(q^R p)^{\alpha-1}} + \frac{1}{2}(pq^{R-1}\frac{1}{1-q^{R-1}})^{\alpha}\right. \\ &\left. + \frac{pq^{-2\alpha+(1-\alpha)R+1}}{2(1-q^{R-1})} \cdot \frac{q^{4\alpha-2}(1-q^{(2\alpha-1)(R-2)})}{1-\ q^{4\alpha-2}}\right\}\end{aligned} \tag{23}$$

where $q = 1-p$. Thus, we have

$$\begin{aligned}\epsilon(\alpha) \geq \frac{1}{\alpha-1}\log\left\{\frac{1}{2}\frac{(1-q^{R-1})^{\alpha-1}}{(q^R p)^{\alpha-1}} + \frac{1}{2}(pq^{R-1}\frac{1}{1-q^{R-1}})^{\alpha}\right. \\ \left. + \frac{pq^{-2\alpha+(1-\alpha)R+1}}{2(1-q^{R-1})} \cdot \frac{q^{4\alpha-2}(1-q^{(2\alpha-1)(R-2)})}{1-\ q^{4\alpha-2}}\right\}.\end{aligned} \tag{24}$$

Fig 4 plots the $\epsilon(\alpha)$ curve obtained with a $\mathcal{Q}_{\text{MGeo}}(\cdot)$ with $R = 8$ and $p = 0.5$.

C. Differential privacy for multidimensional QMGeo

1) ϵ-DP: For any individual agent's gradient, we apply the scalar $Q_{\text{QMGeo}}(\cdot)$ to every element in the gradient. The entire privacy mechanism can be viewed as a function of multiple $Q_{\text{QMGeo}}(\cdot)$ mechanisms. Thus, for a gradient vector $\boldsymbol{g} = (g_1, g_2, ..., g_d)$ the final quantized model update is $(Q_{\text{QMGeo}}(g_1), Q_{\text{QMGeo}}(g_2), ..., Q_{\text{QMGeo}}(g_d))$.

Considering the privacy amplification theorem [17] and the random sampling in Section II, according to the sequential

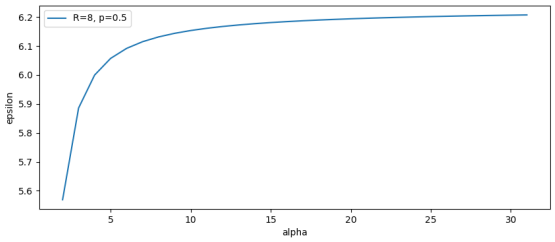

Fig. 4. Plot of $\epsilon(\alpha)$ vs. α. The y-axis is the achieved ϵ, and the x-axis is the corresponding α. The curve is obtained with a $\mathcal{Q}_{\text{MGeo}}(\cdot)$ with $R = 8$ $p = 0.5$.

composition theorem of DP, the overall DP guarantee we obtain for the model update vector with dimension d and sampling rate κ is

$$\epsilon \geq d\kappa(-(\ln p + (R-2)\ln(1-p)) + \ln(1-(1-p)^{R-1})). \tag{25}$$

2) (α, ϵ)-RDP: Similarly, RDP shows a composition of the same pattern [18]. From [19], for $\alpha \leq 2$, the RDP enjoys $\epsilon_{\text{sampled}}(\alpha) = \mathcal{O}(\kappa^2\epsilon(\alpha))$, where κ is the sampling rate. Applying to (24), we have

$$\begin{aligned}\epsilon(\alpha) \geq & \kappa^2 d \frac{1}{\alpha-1} \times \\ & \log\left\{\frac{1}{2}\frac{(1-q^{R-1})^{\alpha-1}}{(q^R p)^{\alpha-1}} + \frac{1}{2}(pq^{R-1}\frac{1}{1-q^{R-1}})^{\alpha}\right. \\ & \left. + \frac{pq^{-2\alpha+(1-\alpha)R+1}}{2(1-q^{R-1})} \cdot \frac{q^{4\alpha-2}(1-q^{(2\alpha-1)(R-2)})}{1-\ q^{4\alpha-2}}\right\}.\end{aligned} \tag{26}$$

V. Optimality Gap

We next characterize how the perturbation introduced by the randomized quantization method impacts the convergence performance by characterizing the optimality gap between $F(\boldsymbol{w}_T)$ (the loss value achieved at global iteration T) and F^*, the optimal loss value.

We introduce two assumptions for this purpose.

a) Assumption 1: L-smoothness is assumed for the loss function F.

b) Assumption 2: (Polyak-Lojosiewicz Inequality) We assume that the loss function $F(\boldsymbol{w})$ satisfies the Polyak-Lojosiewicz (PL) condition:

$$\frac{1}{2}\|\nabla F(\boldsymbol{w})\|^2 \geq \mu[F(\boldsymbol{w}) - F^*]. \tag{27}$$

We abuse the notation of $Q_{\text{MGeo}}(\boldsymbol{g}_t)$ a bit in this section to represent the operation of applying scalar $Q_{\text{MGeo}}(\cdot)$ to every element in the vector. We define $\boldsymbol{\delta}_t = Q_{\text{MGeo}}(\boldsymbol{g}_t) - \nabla F(\boldsymbol{w}_t)$, to characterize the perturbation introduced by the randomized quantization. Following the first steps of [20] and taking $\boldsymbol{\delta}_t$ into account, we have

$$\begin{aligned}F(\boldsymbol{w}_{t+1}) &\leq F(\boldsymbol{w}_t) + \nabla F(\boldsymbol{w}_t)^T(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \\ &+ \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\ &\leq F(\boldsymbol{w}_t) - \nabla F(\boldsymbol{w}_t)^T\eta(\nabla F(\boldsymbol{w}_t) + \boldsymbol{\delta}_t) \\ &+ \eta^2\frac{L}{2}\|\nabla F(\boldsymbol{w}_t) + \boldsymbol{\delta}_t\|^2\end{aligned} \tag{28}$$

$$\begin{aligned}F(\boldsymbol{w}_{t+1}) \leq F(\boldsymbol{w}) &- \eta(1-\frac{\eta L}{2})\|\nabla F(\boldsymbol{w}_t)^T\|^2 \\ &+ \eta^2\frac{L}{2}\|\boldsymbol{\delta}_t\|^2 \\ &+ \eta(-1+\eta L)\nabla F(\boldsymbol{w}_t)^T\boldsymbol{\delta}_t.\end{aligned} \tag{29}$$

From (29), we substract F^* from both sides, which gives us

$$\begin{aligned}&F(\boldsymbol{w}_{t+1}) - F^* \\ &\leq F(\boldsymbol{w}_t) - F^* - \eta(1-\frac{\eta L}{2})\|\nabla F(\boldsymbol{w}_t)^T\|^2 \\ &+ \eta^2\frac{L}{2}\|\boldsymbol{\delta}_t\|^2 + \eta(-1+\eta L)\nabla F(\boldsymbol{w}_t)^T\boldsymbol{\delta}_t.\end{aligned} \tag{30}$$

We next apply the PL condition to (30), which leads to

$$\begin{aligned}&F(\boldsymbol{w}_{t+1}) - F^* \\ &\leq (1-2\mu\eta(1-\frac{\eta L}{2}))(F(\boldsymbol{w}_t) - F^*) + \eta^2\frac{L}{2}\|\boldsymbol{\delta}_t\|^2 \\ &+ \eta(-1+\eta L)\nabla F(\boldsymbol{w}_t)^T\boldsymbol{\delta}_t.\end{aligned} \tag{31}$$

For simplicity, let

$$X = 1 - 2\mu\eta(1-\frac{\eta L}{2}) \tag{32}$$

$$Y = \eta^2\frac{L}{2}\|\boldsymbol{\delta}_t\|^2 \tag{33}$$

$$Z = \eta(-1+\eta L)\nabla F(\boldsymbol{w}_t)^T\boldsymbol{\delta}_t. \tag{34}$$

Applying (31) iteratively gives us

$$F(\boldsymbol{w}_{tot}) - F^* \leq X^{tot}(F(\boldsymbol{w}_0) - F^*) + \sum_{a=1}^{tot-1}(Y+Z)X^{tot-a-1}. \tag{35}$$

We note that for a fixed choice of parameters of $Q_{\text{MGeo}}(\cdot)$, it is obvious that the term $\boldsymbol{\delta_t}$ is bounded in norm. Thus, (35) indicates that with a correct choice of parameters, convergence is guaranteed for the framework.

VI. Numerical Results

The following results are obtained on MNIST dataset (10 class handwritten digit dataset) with a total of 60000 samples. The dataset is distributed as follows: 90% of the data are randomly split into 5 users, and 10% of the data are reserved at the PS for evaluation. The FL framework trains a multi-layer perceptron with one hidden layer of 32 nodes. We first perform principal component analysis on the dataset and reduce the dimensionality of the input data to 100 to speed up

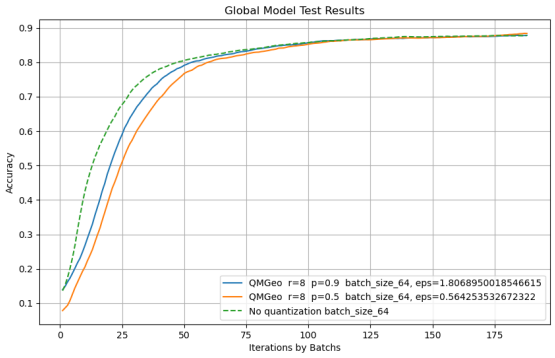

Fig. 5. The two solid line curves displayed in this figure correspond to $p = 0.5$ and $p = 0.9$ in $Q_{\text{QMGeo}}(\cdot)$ respectively, with the number of quantization levels $R = 8$. The dotted curve is the baseline curve where we do not apply any quantization. The y-axis shows the accuracy obtained at the global model using the hold-out test set, while the x-axis shows the number of communication rounds. The ϵ labeled in the legend is acquired using (26), fixing $\alpha = 2$.

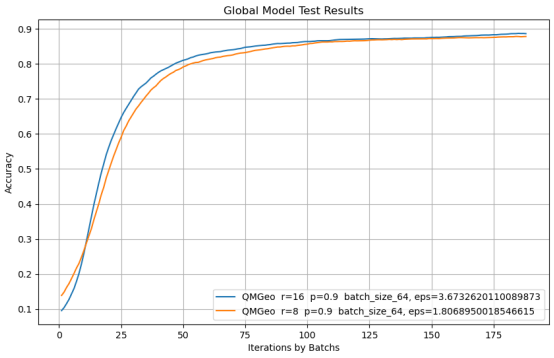

Fig. 6. The two curves displayed in this figure correspond to $R = 16$ and $R = 8$ in $Q_{\text{QMGeo}}(\cdot)$ respectively, with the parameter of success rate $p = 0.9$. The y-axis shows the accuracy obtained at the global model using the hold-out test set, while the x-axis shows the number of communication rounds. The ϵ labeled in the legend is acquired using (26), fixing $\alpha = 2$.

the computation. The dimension of the model update vector is $d = 3562$.

We set batch size to 64 to achieve a sub-sampling rate of $\kappa = 0.005333$, learning rate η to 0.04 and apply a clipping threshold of 0.05 to each element in the gradient vector.

We use (26) to calculate the $\epsilon(\alpha)$ per round achieved by the $Q_{\text{QMGeo}}(\cdot)$ method. Fig 5 demonstrates the accuracy comparison between setting $p = 0.9$ and $p = 0.5$, when the number of quantization levels is fixed at $R = 8$. The dotted line is a baseline case in which no quantization is performed. All curves approach a similar accuracy level, indicating that quantization does not result in performance degradation in terms of accuracy at least in the considered setup. In terms of privacy guarantees, using $Q_{\text{QMGeo}}(\cdot)$ with $p = 0.9$ and $R = 8$ achieves $\epsilon = 1.807$ per round, and $Q_{\text{QMGeo}}(\cdot)$ with $p = 0.5$ and $R = 8$ achieves $\epsilon = 0.564$ per round.

Fig 6 demonstrates the accuracy comparison between setting $R = 16$ and $R = 8$ for fixed parameter $p = 0.9$. Using $Q_{\text{QMGeo}}(\cdot)$ with $p = 0.9$ and $R = 16$ achieves $\epsilon = 3.673$ per round, while $Q_{\text{QMGeo}}(\cdot)$ with $p = 0.9$ and $R = 8$ leads to $\epsilon = 1.807$ per round.

VII. Conclusions

In this paper, we have presented a novel stochastic quantization method $Q_{\text{QMGeo}}(\cdot)$, that utilizes a mixture of truncated geometric distributions to provide randomness for differential privacy. While reducing communication cost through quantization, we also achieve ϵ-DP without the need of any additive noise. In particular, we have provided a privacy analysis for $Q_{\text{QMGeo}}(\cdot)$ both in terms of ϵ-DP and RDP, and demonstrated that certain differential privacy levels can be achieved via properly designed stochastic quantization. We have further conducted an optimality gap analysis that mathematically characterizes the convergence performance of the FL framework.

References

[1] P. Kairouz and et al., "Advances and open problems in federated learning," 2021.

[2] 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*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[3] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," *CoRR*, vol. abs/2006.10672, 2020. [Online]. Available: https://arxiv.org/abs/2006.10672

[4] 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.

[5] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, may 2019. [Online]. Available: https://doi.org/10.1109%2Fsp.2019.00065

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," 2017.

[7] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Towards demystifying membership inference attacks," *ArXiv*, vol. abs/1807.09173, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:50778569

[8] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[9] Ú. Erlingsson, A. Korolova, and V. Pihur, "RAPPOR: randomized aggregatable privacy-preserving ordinal response," *CoRR*, vol. abs/1407.6981, 2014. [Online]. Available: http://arxiv.org/abs/1407.6981

[10] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," *CoRR*, vol. abs/1712.01524, 2017. [Online]. Available: http://arxiv.org/abs/1712.01524

[11] J. Abowd, D. Kifer, S. L. Garfinkel, and A. Machanavajjhala, "Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:237407049

[12] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. Mcmahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," 2018.

[13] P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete gaussian mechanism for federated learning with secure aggregation," 2022.

[14] W.-N. Chen, A. Ozgur, and P. Kairouz, "The poisson binomial mechanism for unbiased federated learning with secure aggregation," in

International Conference on Machine Learning. PMLR, 2022, pp. 3490–3506.

[15] Y. Wang and T. Basar, "Quantization enabled privacy protection in decentralized stochastic optimization," 2022.

[16] T. Olatayo, "Truncated geometric bootstrap method for time series stationary process," *Applied Mathematics*, vol. 2014, 2014.

[17] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," 2018.

[18] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017, pp. 263–275.

[19] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1226–1235.

[20] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.