Killing Two Birds with One Stone: Quantization Achieves Privacy in Distributed Learning

Guangfeng Yan[1] Tan Li[1] Kui Wu[2] Linqi Song[1]
[1] Department of Computer Science, City University of Hong Kong
[2] Department of Computer Science, University of Victoria

***Abstract*—Communication efficiency and privacy protection are two critical issues in distributed machine learning. Existing methods tackle these two issues separately and may have a high implementation complexity that constrains their application in a resource-limited environment. We propose a comprehensive quantization-based solution that could simultaneously achieve communication efficiency and privacy protection, providing new insights into the correlated nature of communication and privacy. Specifically, we demonstrate the effectiveness of our proposed solutions in the distributed stochastic gradient descent (SGD) framework by adding binomial noise to the uniformly quantized gradients to reach the desired differential privacy level but with a minor sacrifice in communication efficiency. We theoretically capture the new trade-offs between communication, privacy, and learning performance.**

***Keywords*-Distributed Learning, Communication Efficiency, Quantization, Privacy**

I. Introduction

Distributed machine learning has recently attracted more attention due to the distributed ways of collecting and processing data, such as federated learning. Communication efficiency and privacy protection are two critical concerns in such an emerging paradigm. Various deep learning models need to be exchanged among distributed computing nodes, which are often subject to the communication bottleneck [1], [2]. Furthermore, exchanging these model parameters or even just gradients may lead to the privacy leakage of local data [3]–[5].

Existing works proposed separate methods to tackle these two issues and made corresponding progress. Communication reduction can be achieved by some compression techniques [6]–[9] while privacy leakage can be constrained to a certain level by add-on noise to disturb the raw updated gradients [10], [11]. While some works [12], [13] try to tackle the two problems together, they simply combine the above techniques, e.g., compressing the perturbed gradient, without offering a theoretical guarantee on the final model performance. Besides, none of them consider the privacy proprieties of compression. We argue that the above two operations are not completely independent. In particular, quantization, one of the representative compression techniques, has inherited privacy properties. Therefore, we can kill two birds with one stone.

In this paper, we propose a new quantization-based solution to achieve communication efficiency and privacy protection simultaneously in a distributed stochastic gradient descent (SGD) framework. Instead of performing extra processing on add-on noise, our solution utilizes the inherent quantization noise to achieve the desired level of privacy. Specifically, our proposed **B**inomial mechanism aided **Q**uantized (BQ) scheme adds appropriately parameterized binomial distributed noise to the quantized gradient, which can be combined together into a unified quantization process, to reach the required privacy with only a slight sacrifice in communication efficiency. We theoretically analyze the privacy leakage of the quantized gradients in a differential privacy (DP) manner: how does removing one sample in the training dataset impact the output of the quantization mechanism?

Using the inherent privacy properties of BQ scheme, we design a novel binomial mechanism-aided quantized SGD (BQ-SGD) algorithm, which can satisfy the required privacy communication budget by adjusting both the quantization level and binomial noise level. We theoretically characterize the trade-offs between communication, privacy, and learning performance under this quantization scheme.

II. Related Work

Several schemes have been proposed to improve the communication efficiency in gradient-based large-scale distributed learning: sparsification [14], sketching [15], quantization [6], less frequent communication [9], [16], or their combinations [7], [8]. In particular, we pay special interest to quantization, which is the basic technique that changes the floating point numbers into fixed point ones. In contrast, other techniques can be used together with quantization in parallel.

Privacy preservation for machine learning has been studied in distributed learning algorithms to prevent privacy leakage from the model exchange. To counter this issue, most existing efforts [17] applied DP mechanism to add controllable noise to the to raw gradients. The trade-off between learning convergence and privacy has been theoretically studied in [18].

Some recent work has considered communication-efficiency and privacy-preserving abilities together in distributed learning [12], [13], [19], however, in a straightforward way via a combination of adding noise and quantizing the noisy gradient. Different from them, the privacy properties of compression have been observed by some early works [20], [21], which illustrate that one can translate the perturbation incurred by the compression method into measurable noise to ensure privacy. However, none of them built a theoretical framework to understand this or theoretically proves the privacy guarantee that quantification itself can provide.

III. Problem Formulation

We consider a distributed learning problem, where N clients collaboratively train a shared model via a central parameter server. The local dataset at client i is denoted as $D^{(i)}$. Our goal is to find a set of global optimal model parameters θ by minimizing the objective function $F : \mathbb{R}^d \to \mathbb{R}$,

$$\min_{\theta\in\mathbb{R}^d} F(\theta) = \sum_{i=1}^{N} p_i \mathbb{E}_{\xi^{(i)}\sim D^{(i)}}[l(\theta;\xi^{(i)})], \quad (1)$$

where p_i [1] is the weight of client i; $l(\theta;\xi^{(i)})$ is the local loss function of the model θ towards one data sample $\xi^{(i)}$; and the expectation is taken with respect to the sampling randomness of data $\xi^{(i)}$.

A standard approach to solve this problem is SGD and its variations, such as momentum, or Adam. Without loss of generality, we describe our framework with the classic setting with one local update (one step of gradient descent) for each iteration. Note that our method also works for these gradient-based variations. At iteration t, each client i first downloads the global model θ_t from server, then randomly selects a batch of samples $B_t^{(i)} \subseteq D^{(i)}$ with size L to compute its local stochastic gradient with model parameter θ_t:

$$\mathbf{g}_t^{(i)} = \mathcal{G}(D^{(i)}) := \frac{1}{L}\sum_{\xi^{(i)}\in B_t^{(i)}: B_t^{(i)}\subseteq D^{(i)}} \nabla l(\theta_t;\xi^{(i)}). \quad (2)$$

Then the server aggregates these gradients and sends the updated model θ_{t+1} back to all clients:

$$\theta_{t+1} = \theta_t - \eta \sum_{i=1}^{N} p_i \mathbf{g}_t^{(i)} \quad (3)$$

where η is the server learning rate. We make the following two common assumptions on the raw gradient $\mathbf{g}_t^{(i)}$ and the objective function $F(\theta)$ [24], [25]:

Assumption 1 (Unbiasness and Bounded Variance of $\mathbf{g}_t^{(i)}$)**.** *The stochastic gradient oracle gives us an independent unbiased estimate $\mathbf{g}_t^{(i)}$ with a bounded variance:*

$$\mathbb{E}_{B_t^{(i)}}[\mathbf{g}_t^{(i)}] = \nabla F(\theta_t),\ \mathbb{E}_{B_t^{(i)}}[\|\mathbf{g}_t^{(i)} - \nabla F(\theta_t)\|^2] \le \frac{\sigma^2}{L}. \quad (4)$$

Assumption 2 (Smoothness)**.** *The objective function $F(\theta)$ is ν-smooth: $\forall\theta,\theta' \in \mathbb{R}^d$, $\|\nabla F(\theta) - \nabla F(\theta')\| \le \nu\|\theta - \theta'\|$.*

Assumption 2 further implies that $\forall\theta,\theta' \in \mathbb{R}^d$, we have

$$F(\theta') \le F(\theta) + \nabla F(\theta)^{\mathrm{T}}(\theta' - \theta) + \frac{\nu}{2}\|\theta' - \theta\|^2. \quad (5)$$

In vanilla distributed SGD, client i directly transmits the raw stochastic gradient $\mathbf{g}_t^{(i)}$ to the server, which may pose a significant communication burden and privacy risk. To cope with the two drawbacks, we propose quantization schemes that convert $\mathbf{g}_t^{(i)}$ to a perturbed version and carefully characterize the inherent privacy guarantee it can provide using a DP-like definition. Based on the proposed quantization scheme, we present a new variant of distributed SGD that can achieve the desired level of privacy and communication constraints.

[1]Note that this work focuses on exploring the impact of server-client communication, rather than the specific server-side aggregation on learning performance. We claim that our proposed method can be used in combination with different aggregation methods, such as [9] [22] or [23].

IV. Binomial Mechanism Aided Quantization

In this section, we propose a three-step quantization mechanism, called Binomial mechanism aided Quantization (BQ) and carefully characterize its inherent privacy guarantee.

A. BQ scheme

Step 1: Norm Clipping We first clip the gradient of each sample $\nabla l(\theta_t;\xi^{(i)})$ of Eq. (2) into l_∞ norm with threshold C:

$$\nabla\hat{l}(\theta_t;\xi^{(i)}) = \frac{\nabla l(\theta_t;\xi^{(i)})}{\max\{1, \|\nabla l(\theta_t;\xi^{(i)})\|_\infty / C\}}. \quad (6)$$

We then calculate the clipped gradient by averaging all $\nabla\hat{l}(\theta_t;\xi^{(i)})$ in a batch :

$$\mathbf{g}_t^{(i),C} = \frac{1}{L}\sum_{\xi^{(i)}\in B_t^{(i)}} \nabla\hat{l}(\theta_t;\xi^{(i)}). \quad (7)$$

Since $\|\nabla\hat{l}(\theta_t;\xi^{(i)})\|_\infty \le C$, referring to the triangular inequality of l_∞ norm, we have $\|\mathbf{g}_t^{(i),C}\|_\infty \le C$. The clipped gradient here limits the impact of a single sample on the whole, and it also plays a role in the privacy analysis.

Step 2: Uniform Quantization We then introduce a commonly used quantization scheme, namely, uniform quantization [6] to quantize the clipped gradients in an element-wise way. In particular, the j-th element of $\mathbf{g}_t^{(i),C}$, denoted as g_j for simplicity, is quantized as

$$\mathcal{Q}_b[g_j] = C \cdot \mathrm{sgn}(g_j) \cdot \rho(g_j, s), \quad (8)$$

where $\mathcal{Q}_b[\cdot]$ is a b-bit quantizer, $\mathrm{sgn}(g_j) = \{+1, -1\}$ is the sign of g_j, s is the given quantization level, and $\rho(g_j, s)$ is an unbiased stochastic function that maps scalar $|g_j|/C$ to one of the values in set $\{0, 1/s, 2/s, \ldots, s/s\}$. For example, for $|g_j|/C \in [l/s, (l+1)/s]$, we have

$$\rho(g_j, s) = \begin{cases} l/s, & \text{w.p. } 1-p, \\ (l+1)/s, & \text{w.p. } p = \frac{s|g_j|}{C} - l. \end{cases} \quad (9)$$

The quantization level s is roughly exponential to the number of quantized bits b. If we use one bit to represent its sign and the other $b-1$ bits to express $\rho(g_j, s)$, we can achieve quantization level $s = 2^{b-1} - 1$. After quantization, each client obtains a tuple $(\boldsymbol{\sigma}_t^{(i)}, \boldsymbol{\rho}_t^{(i)})$, where $\boldsymbol{\sigma}_t^{(i)}$ and $\boldsymbol{\rho}_t^{(i)}$ are the vectors of signs and integer values (i.e., $s\cdot\rho(g_j, s)$) for all elements of $\mathbf{g}_t^{(i),C}$. We define the gradient after uniform quantization as:

$$\mathbf{g}_t^{(i),U} = \frac{C}{s} \cdot \boldsymbol{\sigma}_t^{(i)} \odot \boldsymbol{\rho}_t^{(i)}, \quad (10)$$

where $\odot$ denotes the element-wise product.

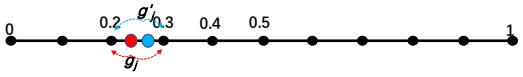
We next show that uniform quantization can provide privacy protection in some cases. In particular, we study the privacy loss of the $\mathbf{g}_t^{(i),U}$ in a DP manner: how does removing or adding one sample in the original dataset impact the output gradient. The formal definition of (ϵ, δ)-DP for SGD is as follows.

Definition 1 ((ϵ, δ) - DP **for SGD).** Given a set of data sets $\mathcal{D}$ and a query function $q : \mathcal{D} \to \mathcal{X}$, a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{O}$ to release the answer of the query, is defined to be (ϵ, δ) - DP if for any adjacent datasets $(D, D') \in \mathcal{D} \times \mathcal{D}$ and any measurable subset outputs $O \in \mathcal{O}$,

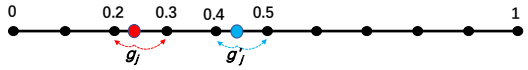
$$\Pr\{\mathcal{M}[q(D)] \in O\} \leq \Pr\{\mathcal{M}[q(D')] \in O\}e^{\epsilon} + \delta, \quad (11)$$

where $\epsilon > 0$ is the distinguishable bound of all outputs on adjacent datasets D, D' that differ in at most one data sample. δ represents the event that the ratio of the probabilities for two adjacent datasets D, D' cannot be bounded by e^{ϵ} after privacy-preserving mechanism $\mathcal{M}$. In the rest of the paper, we also call a solution *SGD private* if it meets the above definition.

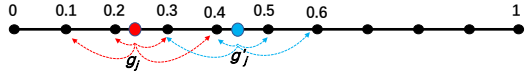
In the SGD framework, $\mathcal{G}(D)$ is viewed as a *query function* that maps dataset D to a *query result* (i.e., raw gradient $\mathbf{g}$). The SGD privacy ensures that any element of the $\mathbf{g}$ is "essentially" equally likely to occur, independent of the presence or absence of any individual data sample. We denote $\mathbf{g}, \mathbf{g}'$ as two raw gradients computed based on two adjacent datasets D, D', and g, g' as two elements in the same entry of $\mathbf{g}, \mathbf{g}'$. By considering uniform quantization as a mechanism $\mathcal{M}$ that maps query results (raw gradients $\mathbf{g}$) to the perturbed outputs (quantized gradient $\mathbf{g}^U$), our criterion for detecting whether uniform quantization satisfies SGD privacy is: whether the original g and g' can be distinguished through the disturbed g^U and g'^U.



(a) Uniform Quantization ($s = 10$)



(b) Uniform Quantization ($s = 10$)



(c) Binomial Mechanism Aided Quantization ($s = 10$, $m = 1$)

Fig. 1. An illustration of quantized gradients after step 2 and step 3.

The uniform quantization maps the raw value to the two nearest quantization points, see Eq. (9) and Fig. 1. In Fig. 1(a), g and g' are within interval $[0.2, 0.3]$. We say the uniform quantization satisfies SGD privacy if g and g' are mapped to the same quantization points, i.e. $g^U = g'^U = 0.2(\text{or } 0.3)$. Fig. 1(b) shows the case that uniform quantization fails to provide SGD privacy. g and g' fall in different intervals. After uniform quantization, $g^U = 0.2$ or 0.3, while $g'^U = 0.4$ or 0.5, which g^U and g'^U can always be distinguished. This example suggests that **Step 1 and Step 2 together cannot achieve SGD privacy.**

Step 3: Binomial Noise Addition To enhance the SGD privacy guarantee, additional noise should be added after step 2. Inspired by [19], we add an extra noise $\mathbf{o}_t^{(i)}$ on the values after uniform quantization, which are i.i.d sampled from a Binomial distribution $Bin(m, q)$, to enhance the privacy guarantee with only a slight sacrifice in communication. Fig. 1(c) shows that the binomial noise addition allows g^U and g'^U to reach the yonder points, like $[0.3, 0.4]$, thus achieving SGD privacy. Formally, the uniform Quantization maps g_j into the values in set $\{-s, 1-s, ..., 0, 1, ..., s-1, s\}$. After binomial noise addition, the set is extended to $\{-s, 1-s, ..., 0, 1, ..., s+m-1, s+m\}$, which includes $2s+m+1$ points. That is, $\log(2s+m+1)$ bits are needed for transmission. The client i sends $\boldsymbol{\sigma}_t^{(i)} \odot \boldsymbol{\rho}_t^{(i)} + \boldsymbol{o}_t^{(i)}$ instead of raw gradient $\mathbf{g}_t^{(i)}$ to the server. To ensure the unbiased estimation of $\mathbf{g}_t^{(i)}$, the server decodes the tuple and computes the BQ gradient as:

$$\mathbf{g}_t^{(i),B} = \mathcal{G}_B(D^{(i)}) = \frac{C}{s} \cdot [\boldsymbol{\sigma}_t^{(i)} \odot \boldsymbol{\rho}_t^{(i)} + \mathbf{o}_t^{(i)} - mq \cdot \mathbf{1}], \quad (12)$$

where $\mathbf{1}$ is an all-one vector.

B. Inherit privacy guarantee of BQ scheme

We then investigate how the BQ gradient $\mathbf{g}_t^{(i),B}$ protects privacy of dataset $D^{(i)}$ being used for model training. Most DP-manner mechanisms, like Laplace or Gaussian mechanism, add controlled noise from some predetermined distributions on raw gradient $\mathbf{g}$ to ensure privacy. Analogously, the BQ gradient $\mathbf{g}_t^{(i),B}$ can be viewed as the noisy version of original one $\mathbf{g}_t^{(i)}$, namely, $\mathbf{g}_t^{(i),B} = \mathbf{g}_t^{(i)} + \mathbf{r}_t^{(i)}$. Therefore, the properties of BQ noise $\mathbf{r}_t^{(i)}$ is the key factor for the privacy analysis. We show the properties of the quantization noise in Proposition 1.

Proposition 1 (Properties of BQ noise)**.** *After performing C-norm clipping, s-level quantization and adding binomial noise vector i.i.d sampled from $Bin(m, q)$, the probability density of the j-th element of $\boldsymbol{r}_t^{(i)}$, denote as r_j, is:*

$$f_{\boldsymbol{r}}(r_j) = \frac{s}{C}[(k+1-mq)P_k + (mq-k)P_{k+1}] + \frac{s^2}{C^2}(P_{k+1} - P_k)r_j, \quad (13)$$

$$r_j \in [\tfrac{(k-mq)C}{s}, \tfrac{(k+1-mq)C}{s}], k = -1, 0, ..., m.$$

where $P_k = \binom{m}{k} q^k (1-q)^{m-k}$, $P_{-1} = P_{m+1} = 0$. The statistic properties of BQ noise are : $\mathbb{E}[\boldsymbol{r}] = \mathbf{0}$ and

$$\mathbb{E}[\|\boldsymbol{r}\|^2] = dC^2\left[\tfrac{mq(1-q)}{s^2} + \tfrac{1}{6s^2}\right] \triangleq dC^2 V(m, q, s) \quad (14)$$

We define $V(m, q, s) \triangleq \frac{mq(1-q)}{s^2} + \frac{1}{6s^2}$ as the BQ noise variance. The proof of Proposition 1 is in Appendix VI-A. Using Proposition 1, we can derive the privacy guarantee provided by BQ quantized gradient as follows.

Lemma 1 (Privacy guarantee of BQ Gradient)**.** *The BQ gradient $\tilde{\mathbf{g}}_t^{(i),B}$ satisfies $(\epsilon^{(i)}, \delta^{(i)})$-SGD privacy with $\epsilon^{(i)} = \frac{8dsLP_{max}}{|D^{(i)}|^2\delta^{(i)}}$, where $P_{max} = max_k\{P_k\}$.*

The proof of Lemma 1 is in Appendix VI-B.

Remark 1. *For general m, $q = \frac{1}{2}$ is the optimal choice as it can minimize P_{max}. According to De Moivre–Laplace theorem, for $m > 10$, P_{max} can be expressed as*

$$P_{max} = \max_k\{\tfrac{1}{\sqrt{2\pi\sigma^2}} \exp \tfrac{-(k-\mu)^2}{2\sigma^2}\} = \tfrac{1}{\sqrt{2\pi\sigma^2}}. \quad (15)$$

where $\sigma^2 = \frac{m}{4}$. *Hence,* $\epsilon^{(i)} = \frac{6.4dsL}{|D^{(i)}|^2\sqrt{m}\delta^{(i)}}$.

Remark 2. *The privacy provided by the BQ gradient is affected by three key factors: (i) the dimension of gradient* d*, since each dimension may reveal privacy; (ii) the ratio of batch sampling* $L/|D^{(i)}|$*. The small ratio leads to the small* ϵ *(i.e. stronger privacy) according to the privacy amplification theorem [26]; (iii) quantization level* s *and binomial noise* m*, the lower quantization level* s *and larger binomial noise* m *contribute to stronger privacy protection. Noting since* s *is always larger than 1 and* $2s + m + 1 < 2^b$*, to provide* (ϵ, δ)*-SGD privacy, the required bits* b *for transmission should be no less than* $\frac{1}{2}\log\frac{6.4dL}{|D^{(i)}|^2} - \frac{1}{2}\log\epsilon\delta$.

C. Algorithm Description

The inherent privacy properties of the BQ scheme are stated in Lemma 1. However, many real-world applications will give the privacy requirement and communication budget in advance by considering the sensitivity of local datasets, channel conditions, and tolerant latency. In this section, we propose the following algorithm, Binomial Mechanism aided Quantized distributed SGD (BQ-SGD), where each client takes explicit communication constraint $\bar{b}^{(i)}$ and privacy requirement.

Algorithm 1 Binomial Mechanism aided Quantized distributed SGD (BQ-SGD)

1: **Input:** Learning rate η, initial point $\theta_0 \in \mathbb{R}^d$, gradient norm bound C; set communication constraint $\bar{b}^{(i)}$ and privacy requirement $(\bar{\epsilon}^{(i)}, \bar{\delta}^{(i)})$ for client i.
2: **for** each iteration $t = 0, 1, ..., T-1$: **do**
3: **On each client** $i = 1, ..., N$**:**
4: Receive θ_t from server;
5: Compute $\nabla l(\theta_t; \xi^{(i)})$ for each sample $\xi^{(i)} \in B_t^{(i)}$;
6: Clipping $\nabla l(\theta_t; \xi^{(i)})$ to $\nabla \hat{l}(\theta_t; \xi^{(i)})$ using Eq. (6);
7: Determine $(s^{(i)}, m^{(i)})$ for client i using Eqs. (19) (20);
8: Compute parameters $\boldsymbol{\sigma}_t^{(i)}, \boldsymbol{\rho}_t^{(i)}$ for uniform quantizer, and $\boldsymbol{o}_t^{(i)}$, where $\mathbf{o}_t^{(i)}$ are i.i.d. samples from $Bin(m^{(i)}, \frac{1}{2})$;
9: Send $(\boldsymbol{\sigma}_t^{(i)} \odot \boldsymbol{\rho}_t^{(i)} + \boldsymbol{o}_t^{(i)})$ to the server;
10: **On the server:**
11: Decode $\mathbf{g}_t^{(i),B}$ according to Eq. (12);
12: Aggregate $\bar{\mathbf{g}}_t \triangleq \sum_{i=1}^N p_i \mathbf{g}_t^{(i),B}$;
13: Update model parameter: $\theta_{t+1} = \theta_t - \eta\bar{\mathbf{g}}_t$;
14: Send θ_{t+1} to all clients;
15: **end for**

The key step of Alg. 1 is Line 7, which determines the quantization level $s^{(i)}$ and binomial noise level $m^{(i)}$ to meet the needs of $\bar{b}^{(i)}$ and $(\bar{\epsilon}^{(i)}, \bar{\delta}^{(i)})$ for each client. Note that if $\bar{b}^{(i)}$ and $(\bar{\epsilon}^{(i)}, \bar{\delta}^{(i)})$ are invariant during T iterations, we only need to calculate $s^{(i)}$ and $m^{(i)}$ once before training starts. But if the requirements change, we need to recalculate the parameters for the BQ scheme. For simplicity, we omit i and t in the following analysis. We seek a pair of (s, m) that minimize the convergence error under privacy and communication constraints, which is further equivalent to minimizing the BQ noise variance $V(m, q, s)$ defined in Eq. (14). By setting $q = \frac{1}{2}$, we can formally formulate this constrained optimization problem as:

$$\min_{m,s} V(m, \frac{1}{2}, s) = \frac{m}{4s^2} + \frac{1}{6s^2} \tag{16}$$

$$s.t \quad \log(2s + m + 1) \le \bar{b}, \tag{17}$$

$$\frac{6.4dsL}{|D|^2\sqrt{m}\bar{\delta}} = \bar{\epsilon}. \tag{18}$$

Eq. (17) captures the communication constraint using $2s + m < 2^{\bar{b}} - 1$. Eq. (18) indicates the privacy constraint, which can be directly derived from Lemma 1. By solving the above optimization problem, we have,

$$s = \bar{R}\sqrt{\bar{R}^2 + (2^{\bar{b}} - 1)} - \bar{R}^2, \tag{19}$$

$$m = \frac{s^2}{\bar{R}^2} = (2^{\bar{b}} - 1) + 2\bar{R}^2 - 2\bar{R}\sqrt{\bar{R}^2 + (2^{\bar{b}} - 1)} \tag{20}$$

where $\bar{R} = \frac{\bar{\delta}\bar{\epsilon}|D|^2}{6.4dL}$ is proportional to the privacy level.

D. Performance Analysis

Theorem 1 (Performance of Algorithm 1)**.** *For an* N*-client distributed learning problem in Eq.* (1)*, given the communication constraint* $\bar{b}^{(i)}$*, privacy level* $(\bar{\epsilon}^{(i)}, \bar{\delta}^{(i)})$*, clipping norm bound* C*, learning rate* $\eta \le \frac{1}{\nu}$ *and training iterating* T*, BQ-SGD satisfies the following.*

Privacy: *BQ-SGD is* $(\sqrt{2T\log\frac{1}{\bar{\delta}^{(i)}}}\bar{\epsilon}^{(i)}, T\bar{\delta}^{(i)})$*- SGD privacy for client* i*;*

Communication: *BQ-SGD incurs communication cost* $Td\bar{b}^{(i)}$ *for client* i*;*

Convergence: *The convergence error for smooth objectives* $F(\theta_t)$ *is upper bounded by*

$$\frac{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_t)\|^2]}{\le \underbrace{\frac{2[F(\theta_0) - F(\theta^*)]}{T\eta} + \frac{N\sigma^2\sum_{i=1}^N p_i^2}{L}}_{\text{Error of Distributed SGD}} + \underbrace{NdC^2\sum_{i=1}^N p_i^2 V(m^{(i)}, \frac{1}{2}, s^{(i)})}_{\text{Quantization Error}},}$$

where m *and* s *are determined by Eqs.* (19) *and* (20)*.*

The communication and privacy performance can be directly derived from the training procedure, and the full proof of the convergence error bound is shown in Appendix VI-C.

- **Communication**. At each iteration, client i quantizes each element of the d-dimensional gradient from b_{init} to $\bar{b}^{(i)}$ bits (b_{init} is the number of bits of full-precision floating point, e.g., $b_{init} = 32$ or $b_{init} = 64$). We can achieve this communication cost by summing over T iterations. Compared with vanilla distributed SGD, we can reduce $Td(b_{init} - \bar{b}^{(i)})$ bits communication overhead for client i.
- **Privacy**. Using the following strong composition Theorem [10], client i has a total $(\sqrt{T\log\frac{1}{\bar{\delta}^{(i)}}}\bar{\epsilon}^{(i)}, T\bar{\delta}^{(i)})$- SGD privacy after repeating T times BQ operation.

Lemma 2 (Strong Composition Theorem [10]). *For all ϵ, δ $\delta' \geq 0$, the class of (ϵ, δ)- DP mechanisms satisfies $(\epsilon^T, T\delta + \delta')$-differential privacy under T-fold adaptive composition for:*

$$\epsilon^T = \sqrt{-2T \ln \delta'}\epsilon + T\epsilon(e^\epsilon - 1). \tag{21}$$

For small ϵ, $e^\epsilon - 1 \to 0$, we have $\epsilon^T = O(\sqrt{T}\epsilon)$.

• **Trade-off among privacy-communication-convergence** We then fix one of $\bar{\epsilon}$ and $\bar{b}$ to see how the change of the other affects the values of m and s.

Fix privacy constraint $\bar{\epsilon}$. As $\bar{R}$ is proportional to $\bar{\epsilon}$, we turn to analyze the effect of $\bar{b}$ with fixed $\bar{R}$. Using Eq. (20), we rewrite the BQ noise variance $V(m, \frac{1}{2}, s)$ as:

$$V(m, \frac{1}{2}, s) = \frac{1}{4\bar{R}^2} + \frac{1}{6s^2}. \tag{22}$$

For a small communication budget $\bar{b}$, we have to quantize the raw gradient with lower-level $2^{\bar{b}}-1$, which leads to lower-level s according to Eq. (19). For fixed $\bar{R}$ and lower s, the variance tends to be large, indicating that small-level quantization leads to large performance degradation.

Fix communication constraint $\bar{b}$. For high privacy requirements $\bar{R}$, again using Eq. (20) and $m + 2s = 2^{\bar{b}} - 1$, we can obtain a high-level m and low-level s. This demonstrates that the algorithm needs to quantize the gradient to a small level to bring more noise and prefers to add more binomial noise as the additional source to complete $\bar{R}$ to achieve a high privacy guarantee. Taking such values of s and m, the variance function $V(m, \frac{1}{2}, s)$ tends to be large, which means high-level privacy leads to a large performance decay.

V. Experiments

In this section, we conduct experiments on MNIST and Fashion-MNIST to empirically validate our proposed BQ-SGD. We use LeNet-5 [27] for MNIST, and AlexNet [28] for the Fashion-MNIST for all clients. More experimental details are given in Table I.

TABLE I
Experiment Setting.

Dataset	MINIST	Fashion-MNIST
Net	LeNet-5	AlexNet
Dim for SGD Privacy[1]	$d_P = 3 \times 10^3$	$d_P = 3 \times 10^4$
Clipping Threshold	$C = 0.0015$	$C = 0.003$
Learning Rate	0.006	0.006
Batch Size	32	32
Number of Clients	4	4
Size of Local Datasets	15000	15000
Iterations	1000	3000

Performance of BQ-SGD for different privacy constraints. We first fix the communication budget $\bar{b}$ and examine the performance of BQ-SGD under various privacy demands. We set $\bar{b} = 8$ bits, $\bar{\delta} = 1 \times 10^{-4}$ for MNIST and $b^* = 10$ bits, $\bar{\delta} = 1 \times 10^{-4}$ for Fashion-MNIST. From Table II, our BQ-SGD achieves the accuracy of 94.20%, 96.73% and 97.38% on the MNIST datasets with $\bar{\epsilon} = 1.72, 3.44, 8.72$. That is, the model performance deteriorates with the decrease in $\bar{\epsilon}$, indicating stronger privacy requirements correspond to worse model performance. Similar results can be found on Fashion-MNIST, our BQ-SGD achieves the accuracy of 79.11%, 84.16% and 87.02% with $\bar{\epsilon} = 86.22, 112.42, 138.79$. In addition, when faced high privacy demand, i.e., small $\bar{\epsilon}$, we have small s, i.e., low quantization level and large m, i.e., high Binomial noise level since they can incur more noise to meet the requirements of privacy, which is consistent with our conclusion in Theorem 1.

[1]we use $d_P = \|\nabla l\|_1/C$ instead of d to give an approximate value of the privacy level.

TABLE II
BQ-SGD with fixed $\bar{b}$ and different $\bar{\epsilon}$($\bar{\delta} = 1 \times 10^{-4}$)

Dataset	$\bar{\epsilon}$	BQ parameters	Accuracy
MNIST ($\bar{b} = 8$ bits)	1.72	$s = 1, m = 251$	94.20%
	3.44	$s = 2, m = 251$	96.73%
	8.72	$s = 4, m = 247$	97.38%
Fashion-MNIST ($\bar{b} = 10$ bits)	86.22	$s = 10, m = 1003$	79.11%
	112.42	$s = 13, m = 997$	84.16%
	138.79	$s = 16, m = 991$	87.02%

Performance of BQ-SGD for different communication budgets. We next fix the privacy constraint and explore the effect on communication budget $\bar{b}$. We set $\bar{\epsilon} = 3.44$, $\bar{\delta} = 10^{-4}$ for MNIST and $\bar{\epsilon} = 112.42$, $\bar{\delta} = 1 \times 10^{-4}$ for Fashion-MNIST. Table III demonstrates that as the communication budget increases, the performance of our BQ-SGD improves. In particular, on MNIST, the BQ-SGD achieves the accuracy of 96.73%, 96.91%, and 96.99% using $\bar{b} = 8, 10, 12$ bits. And on Fashion-MNIST, the BQ-SGD achieves the accuracy of 84.16%, 84.71%, and 85.02% using $\bar{b} = 10, 12, 14$ bits Furthermore, it can be seen that both s and m grow when the communication constraints are relaxed. On the one hand, clients are allowed to use a high quantization level to ensure learning performance. On the other hand, a larger s incur less disturbance to the raw gradient, thus a larger binomial noise needs to be added to ensure privacy.

TABLE III
BQ-SGD with fixed $\bar{\epsilon}$ and different $\bar{b}$ ($\bar{\delta} = 1 \times 10^{-4}$).

Dataset	$\bar{b}$(bits)	BQ parameters	Accuracy
MNIST ($\bar{\epsilon} = 3.44$)	8	$s = 2, m = 251$	96.73%
	10	$s = 4, m = 1050$	96.91%
	12	$s = 6, m = 4079$	96.99%
Fashion-MNIST ($\bar{\epsilon} = 112.42$)	10	$s = 13, m = 997$	84.16%
	12	$s = 26, m = 4043$	84.71%
	14	$s = 52, m = 16279$	85.02%

VI. Conclusion

We provided new insights into the correlated nature of quantization and privacy. Using the inherent privacy properties, we proposed a new quantization-based solution to achieve communication efficiency and privacy protection simultaneously. We demonstrated the effectiveness of our proposed solutions in the distributed SGD framework and theoretically characterized the new trade-offs among communication, privacy, and learning performance.

REFERENCES

[1] Z. Tang, S. Shi, X. Chu, W. Wang, and B. Li, "Communication-efficient distributed deep learning: A comprehensive survey," *arXiv preprint arXiv:2003.06307*, 2020.

[2] Z. Tao and Q. Li, "esgd: Communication efficient distributed deep learning on the edge," in {*USENIX*} *Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[3] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated learning*. Springer, 2020, pp. 17–31.

[4] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[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, 2019, pp. 739–753.

[6] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1709–1720, 2017.

[7] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations," *arXiv preprint arXiv:1906.02367*, 2019.

[8] G. Nadiradze, A. Sabour, P. Davies, S. Li, and D. Alistarh, "Asynchronous decentralized sgd with quantized and local updates," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[10] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[11] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[12] H. Zong, Q. Wang, X. Liu, Y. Li, and Y. Shao, "Communication reducing quantization for federated learning with local differential privacy mechanism," in *2021 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2021, pp. 75–80.

[13] N. Mohammadi, J. Bai, Q. Fan, Y. Song, Y. Yi, and L. Liu, "Differential privacy meets federated learning under communication constraints," *arXiv preprint arXiv:2101.12240*, 2021.

[14] S. Shi, Q. Wang, K. Zhao, Z. Tang, Y. Wang, X. Huang, and X. Chu, "A distributed synchronous sgd algorithm with global top-k sparsification for low bandwidth networks," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 2238–2247.

[15] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "Fetchsgd: Communication-efficient federated learning with sketching," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8253–8265.

[16] J. Zhang and O. Simeone, "Lagc: Lazily aggregated gradient coding for straggler-tolerant and communication-efficient distributed learning," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 962–974, 2020.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[18] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[19] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. McMahan, "cpsgd: communication-efficient and differentially-private distributed sgd," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7575–7586.

[20] S. Xiong, A. D. Sarwate, and N. B. Mandayam, "Randomized requantization with local differential privacy," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2189–2193.

[21] J. A. Jonkman, T. Sherson, and R. Heusdens, "Quantisation effects in distributed optimisation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3649–3653.

[22] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 118–128.

[23] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[24] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[25] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6155–6165.

[26] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[27] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[29] B. Eva, "Principles of indifference," *The Journal of Philosophy*, vol. 116, no. 7, pp. 390–411, 2019.

APPENDIX

A. Proof of Proposition 1

Proof: The noise is raised from two sources, one is the disturb caused by uniform quantization and another is from added Binomial noise. For simplicity, we omit (i) and t in the following analysis. We first decompose the BQ noise as $\boldsymbol{r}^B = \boldsymbol{r}^U + \frac{C}{s}\mathbf{o} - \frac{Cmq}{s}\cdot\mathbf{1}$, where $\boldsymbol{r}^U$ is incurred by the uniform quantizer and $\frac{C}{s}\mathbf{o} - \frac{Cmq}{s}\cdot\mathbf{1}$ is generated from Binomial noise. We then derive the analysis in two steps.

Step 1: Properties of Uniform quantization noise We first show the statistic properties of the uniform quantization noise $\mathbf{r}^U$. Considering $\mathcal{Q}_b[g_j] = C\cdot\text{sgn}(g_j)\cdot\rho(g_j,s)$, then we define $r_j^U = g_j - \mathcal{Q}_b[g_j]$, i.e., the noise added on the clipped gradient after uniform quantization.

To analyze r_j^U, we introduce an intermediate variable $\varphi_j \triangleq \frac{|g_j|}{C} - \rho(g_j,s)$. Denote $\bar{f}_g(x) = \sum_{l=0}^{s-1} f_g(\frac{l}{s}+x), x\in[0,\frac{1}{s})$ as the PDF of $|g|_j/C$ in a single quantization bin. Then for $0<\varphi_j<\frac{1}{s}$, the probability density of φ_j is

$$\begin{aligned} f_{\varphi_j}\{\varphi_j\} &= \sum_{l=0}^{s-1} f_g\Big\{\frac{|g_j|}{C} = \frac{l}{s}+\varphi_j\Big\}\cdot\Delta \\ &\quad\times \Pr\Big\{\rho(g_j,s) = \frac{l}{s}\Big|\frac{|g_j|}{C} = \frac{l}{s}+\varphi_j\Big\}/\Delta \\ &= \sum_{l=0}^{s-1} f_g\Big\{\frac{|g_j|}{C} = \frac{l}{s}+\varphi_j\Big\}\cdot\frac{\frac{1}{s}-\varphi_j}{\frac{1}{s}} \\ &\triangleq \bar{f}_g(\varphi_j)[1-s\varphi_j]. \end{aligned}$$

where $\bar{f}_g(\varphi_j) \triangleq \sum_{l=0}^{s-1} f_g\Big\{\frac{|g_j|}{C} = \frac{l}{s}+\varphi_j\Big\}$, and Δ is a small region containing the point $\frac{|g_j|}{C} = \frac{l}{s}+\varphi_j$.

Similarly, for $-\frac{1}{s}<\varphi_j<0$, we have:

$$\begin{aligned} f_{\varphi_j}\{\varphi_j\} &= \sum_{l=0}^{s-1} f_g\Big\{\frac{|g_j|}{C} = \frac{l+1}{s}+\varphi_j\Big\}\cdot\Delta \\ &\quad\times \Pr\Big\{\rho(g_j,s) = \frac{l+1}{s}\Big|\frac{|g_j|}{C} = \frac{l+1}{s}+\varphi_j\Big\}/\Delta \\ &= \sum_{l=0}^{s-1} f_g\Big\{\frac{|g_j|}{C} = \frac{l+1}{s}+\varphi_j\Big\}\cdot\frac{\frac{1}{s}+\varphi_j}{\frac{1}{s}} \\ &\triangleq \bar{f}_g(\frac{1}{s}+\varphi_j)[1+s\varphi_j]. \end{aligned}$$

Note we have no prior knowledge of the distribution of g_j due to the randomness of gradient calculation at each iteration. Such randomness comes from optimization algorithms (e.g., model initialization, batch selection) or hardware defaults. Based on the principle of indifference [29], the best that we can assume that the normalized gradient $\frac{|g_j|}{C}$ is uniformly distributed in each quantization bin. Under this assumption, we have $\bar{f}_g(x) = s$ and thus,

$$f_{\varphi_j}\{\varphi_j\} = s - s^2|\varphi_j|, \quad \varphi_j\in[-\frac{1}{s},\frac{1}{s}] \tag{23}$$

Considering that $r_j^U = C\cdot\text{sgn}(g_j)\cdot\varphi_j$ and $f_{\varphi_j}\{\varphi_j\}$ is even function, we have probability density of r_j^U is

$$f_{\mathbf{r}}(r_j^U) = \frac{s}{C} - \frac{s^2}{C^2}|r_j^U|, \quad r_j^U\in[-\frac{C}{s},\frac{C}{s}] \tag{24}$$

Step2: properties of BQ noise Recall that $\boldsymbol{r}^B = \boldsymbol{r}^U + \frac{C}{s}\mathbf{o} - \frac{Cmq}{s}\cdot\mathbf{1}$. Define a temporary vector $\boldsymbol{\psi} = \boldsymbol{r}^U + \frac{C}{s}\mathbf{o}$. For $\psi_j\in[\frac{kC}{s},\frac{(k+1)C}{s}]$, the probability density of ψ_j is

$$\begin{aligned} &f_{\boldsymbol{\psi}}\{\psi_j\} \\ &= \Big[f_{\mathbf{r}}\Big\{r_j^U = \psi_j - \frac{(k+1)C}{s}\Big\}\cdot\Delta\cdot P_{k+1} \\ &\quad + f_{\mathbf{r}}\{r_j^U = \psi_j - \frac{kC}{s}\}\cdot\Delta\cdot P_k\Big]/\Delta \\ &= \Big[\frac{s}{C} - \frac{s^2}{C^2}(\frac{(k+1)C}{s} - \psi_j)\Big]P_{k+1} \\ &\quad + \Big[\frac{s}{C} - \frac{s^2}{C^2}(\psi_j - \frac{kC}{s})\Big]P_k \\ &= \frac{s}{C}[(k+1)P_k - kP_{k+1}] + \frac{s^2}{C^2}(P_{k+1}-P_k)\psi_j, \end{aligned}$$

for $k=-1,0,...,m$, where Δ is a small region for r_j^U, and $P_k = \binom{m}{k}q^k(1-q)^{m-k}$ is the probability value of $\mathbf{o}$ at k. We finally derive the probability density function of the BQ noise $r_j = \psi_j - \frac{Cmq}{s}$ as

$$\begin{aligned} &f_{\boldsymbol{r}}\{r_j\} \\ &= \frac{s}{C}[(k+1)P_k - kP_{k+1}] + \frac{s^2}{C^2}(P_{k+1}-P_k)(r_j + \frac{Cmq}{s}) \\ &= \frac{s[(k+1-mq)P_k + (mq-k)P_{k+1}]}{C} + \frac{s^2(P_{k+1}-P_k)r_j}{C^2} \end{aligned}$$

for $r_j\in[\frac{(k-mq)C}{s},\frac{(k+1-mq)C}{s}]$.

■

B. Proof of Lemma 1

Proof: We abbreviate Eq. (13) as $f_{\boldsymbol{r}}(r_j) \triangleq a_k + b_k r_j$. Let $\mathbf{g},\mathbf{g}'$ be clipped gradients and $\mathbf{g}^B,\mathbf{g}^{B'}$ be the corresponding BQ gradients. Using the definition of DP, we are looking at the difference for the same output $\mathbf{y}$,

$$\begin{aligned} &|\ln\frac{\Pr(\mathbf{g}^{B'}=\mathbf{y})}{\Pr(\mathbf{g}^B=\mathbf{y})}| = \Big|\ln\Big\{\prod_{j=1}^{d}\frac{a_{k'_j}+b_{k'_j}(t_j-g'_j)}{a_{k_j}+b_{k_j}(t_j-g_j)}\Big\}\Big| \\ &\le \sum_{j=1}^{d}\Big|\ln\Big\{\frac{a_{k'_j}+b_{k'_j}(t_j-g'_j)}{a_{k_j}+b_{k_j}(t_j-g_j)}\Big\}\Big| \\ &\overset{(a)}{\le} \sum_{j=1}^{d}\ln\{1+\frac{(|b_{k_j}|+|b_{k_j+1}|)\cdot|g'_j-g_j|}{a_{k_j}+b_{k_j}r_j}\}, \end{aligned}$$

where (a) considers g_j and g'_j are located at adjacent quantization bin (i.e., $|g'_j-g_j| < \frac{C}{s}$) if we take $L > 2s$. (Without losing generality, suppose that $k'_j > k_j$). Let $z_j =$

$\ln\{1+\frac{(|b_{k_j}|+|b_{k_j+1}|)\cdot|g'_j-g_j|}{a_{k_j}+b_{k_j}\zeta_j}\}$, then

$$\begin{aligned}
\mathbb{E}[z_j] &= \sum_{k_j=-1}^{m} \int_{\frac{(k_j-mq)C}{s}}^{\frac{(k_j+1-mq)C}{s}} z_j \Big[a_{k_j}+b_{k_j}\zeta_j\Big] d\zeta_j \\
&\le \sum_{k_j=-1}^{m} \int_{\frac{(k_j-mq)C}{s}}^{\frac{(k_j+1-mq)C}{s}} \frac{(|b_{k_j}|+|b_{k_j+1}|)\cdot|g'_j-g_j|}{a_{k_j}+b_{k_j}\zeta_j} \\
&\qquad \times [a_{k_j}+b_{k_j}\zeta_j] d\zeta_j \\
&= \frac{C\cdot|g'_j-g_j|\sum_{k_j=-1}^{m}(|b_{k_j}|+|b_{k_j+1}|)}{s} \\
&= \frac{2s\cdot|g'_j-g_j|\sum_{k_j=-1}^{m}|P_{k_j+1}-P_{k_j}|}{C} \\
&\le \frac{4s\cdot|g'_j-g_j|P_{max}}{C},
\end{aligned}$$

where $P_{max}=\max_k\{P_k\}$. Hence,

$$\mathbb{E}[\sum_{j=1}^{d} z_j] = \sum_{j=1}^{d}\mathbb{E}[z_j] \le \frac{4s\Delta\mathbf{g}P_{max}}{C}$$

where $\Delta\mathbf{g} = \max_{D,D'}\|\mathbf{g}-\mathbf{g}'\|_1$ is the sensitivity function. According to the Markov's inequality, we have $\Pr\Big[\sum_{j=1}^{d} z_j \ge \epsilon\Big] \le \frac{4s\Delta\mathbf{g}P_{max}}{C\epsilon'} = \delta'$. Note that $\Delta\mathbf{g}$ gives an upper bound on how much we must perturb the output to preserve privacy, therefore it only relates to the norm clipping operation with threshold C. Using Eq. (7), we can rewrite the sensitivity function as:

$$\Delta\mathbf{g} = \frac{1}{L}\max_{\xi,\xi'}\|\nabla\hat{l}(\xi)-\nabla\hat{l}(\xi')\|_1 \le \frac{2dC}{L}\ , \tag{25}$$

where ξ and ξ' are the single entries which are different for datasets D and D'. Leveraging this bound, we have $\epsilon' = \frac{8dsP_{max}}{L\delta'}$. In addition, $\tilde{\mathbf{g}}_t^{(i),B}$ is computed over batch $B_t^{(i)}$, which is randomly sampled from dataset $D^{(i)}$. therefore, according to the privacy amplification theorem [26], $\tilde{\mathbf{g}}_t^{(i)}$ at client i assures $(\frac{L}{|D^{(i)}|}\epsilon', \frac{L}{|D^{(i)}|}\delta')$-differential privacy. Hence, the BQ gradient satisfied $(\epsilon^{(i)},\delta^{(i)})$-differential privacy for client i in one communication round, where $\epsilon^{(i)} = \frac{8dsLP_{max}}{|D^{(i)}|^2\delta^{(i)}}$. ■

C. Proof of Theorem 1

Proof: Combining Assumption 1 and Eq. (14) in Proposition 1, the properties of aggregated gradient $\bar{\mathbf{g}}_t$ at server satisfy:

$$\mathbb{E}_{B_i,\mathcal{Q}}[\bar{\mathbf{g}}_t] = \nabla F(\theta_t), \tag{26}$$

$$\begin{aligned}
\mathbb{E}_{B_i,\mathcal{Q}}\left[||\bar{\mathbf{g}}_t||^2\right] &\le \|\nabla F(\mathbf{x}_t)\|^2 + \frac{N\sigma^2\sum_{i=1}^{N}p_i^2}{L} \\
&\quad + dNC^2\sum_{i=1}^{N}p_i^2 V(m^{(i)},q,s^{(i)}).
\end{aligned} \tag{27}$$

Firstly, we consider function F is ν-smooth, and use Eq. (5):

$$F(\theta_{t+1}) \le F(\theta_t) + \nabla F(\theta_t)^{\mathrm{T}}(\theta_{t+1}-\theta_t) + \frac{\nu}{2}\|\theta_{t+1}-\theta_t\|^2.$$

For the BQ-SGD, $\theta_{t+1} = \theta_t - \eta\bar{\mathbf{g}}_t$, so:

$$F(\theta_{t+1}) \le F(\theta_t) + \nabla F(\theta_t)^{\mathrm{T}}(-\eta\bar{\mathbf{g}}_t) + \frac{\nu\eta^2}{2}\|\bar{\mathbf{g}}_t\|^2.$$

Taking total expectations and using Eq. (26) and (27):

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1})] &\le F(\theta_t) + (-\eta+\frac{\nu\eta^2}{2})\|\nabla F(\theta_t)\|^2 \\
&\quad + \frac{\nu\eta^2 N\sigma^2\sum_{i=1}^{N}p_i^2}{2L} + \frac{\nu\eta^2 dNC^2\sum_{i=1}^{N}p_i^2 V(m^{(i)},q,s^{(i)})}{2}.
\end{aligned}$$

Subtracting $F(\theta_t)$ from both sides, and for $\eta \le \frac{1}{\nu}$

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1})] - F(\theta_t) &\le -\frac{\eta}{2}\|\nabla F(\theta_t)\|^2 + \frac{N\eta\sigma^2\sum_{i=1}^{N}p_i^2}{2L} \\
&\quad + \frac{N\eta dC^2\sum_{i=1}^{N}p_i^2 V(m^{(i)},q,s^{(i)})}{2}.
\end{aligned}$$

Applying it recursively, this yields:

$$\begin{aligned}
\mathbb{E}[F(\theta_T)] - F(\theta_0) &\le -\frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla F(\theta_t)\|^2 + \frac{\eta N\sigma^2 T\sum_{i=1}^{N}p_i^2}{2L} \\
&\quad + \frac{\eta NdTC^2\sum_{i=1}^{N}p_i^2 V(m^{(i)},q,s^{(i)})}{2}.
\end{aligned}$$

Considering that $F(\theta_T) \ge F(\theta^*)$, so:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_t)\|^2] &\le \frac{2[F(\theta_0)-F(\theta^*)]}{T\eta} \\
&\quad + \frac{N\sigma^2\sum_{i=1}^{N}p_i^2}{L} + NdC^2\sum_{i=1}^{N}p_i^2 V(m^{(i)},q,s^{(i)}).
\end{aligned}$$

■