Truncated Non-Uniform Quantization for Distributed SGD

Guangfeng Yan[a], Tan Li[b], Yuanzhang Xiao[c], Congduan Li[d] and Linqi Song[a]

[a]Department of Computer Science, City University of Hong Kong, Hong Kong SAR
[b]Department of Computer Science, The Hang Seng University of Hong Kong, Hong Kong SAR
[c]Department of Electrical and Computer Engineering, University of Hawaii at Manoa, United States
[d]School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China

***Abstract*—In distributed machine learning, communication overhead is a critical bottleneck that hinders scalability and efficiency. Addressing this challenge, our work introduces a novel two-stage quantization strategy designed to enhance the communication efficiency of distributed Stochastic Gradient Descent (SGD). The proposed method initially employs truncation to mitigate the impact of long-tail noise, followed by a non-uniform quantization of the post-truncation gradients based on their statistical characteristics. We provide a comprehensive convergence analysis of the quantized distributed SGD, establishing theoretical guarantees for its performance. Furthermore, by minimizing the convergence error, we derive optimal closed-form solutions for the truncation threshold and non-uniform quantization levels under given communication constraints. Both theoretical insights and extensive experimental evaluations demonstrate that our proposed algorithm outperforms existing quantization schemes, striking a superior balance between communication efficiency and convergence performance.**

***Index Terms*—Distributed Learning, Communication Efficiency, Non-Uniform Quantization, Gradient Truncation, Convergence Analysis**

I. Introduction

The advent of distributed machine learning has been transformative for a wide field of applications, including the Internet of Things (IoT) and federated learning. By decentralizing the learning process, distributed machine learning enables computation to be performed closer to where data is generated, thereby reducing latency, preserving privacy, and leveraging the collective power of edge devices. Distributed Stochastic Gradient Descent (DSGD) [1], [2] is a popular algorithm that utilizes local client data to build distributed models. In DSGD, local gradients are computed at each client and transferred to a parameter server, which aggregates them to update the global model. This process is repeated until all nodes reach a global consensus on the learning model. DSGD is a powerful tool for distributed machine learning, enabling efficient and scalable training of models across a network of devices.

However, the frequent exchange of local gradients can strain limited communication resources, posing a significant challenge to the efficiency and practicality of distributed learning frameworks. To alleviate this bottleneck, various model compression techniques have been applied on local gradients to enhance communication efficiency. Sparsification and quantization are among the most widely adopted strategies.

Among the prevalent quantization methods, uniform quantizers [3] are commonly utilized due to their simplicity and ease of implementation. However, this approach does not adequately represent the typical bell-shaped with a long-tailed distribution of weights and activations in neural networks [4]. A natural approach to managing gradients with a long-tail distribution is to implement gradient truncation [5], [6], which involves establishing a threshold that serves to mitigate the impact of extreme gradient values on quantization. Another method is to design a non-uniform quantizer [7], [8]. Most of the work in these two areas largely stems from empirical engineering practices, such as manually setting clipping thresholds or assigning more quantization points to areas of high data density—the peaks of the distribution—and fewer points within the less dense tails.

Only a handful of studies attempt to theoretically guide the designing of quantizers, achieving limited success. For example, [5] assumes a Laplace distribution for the gradients to find the optimal truncation threshold, but then applies a simple uniform quantizer for the compression; while another study [9], introduces a non-uniform quantizer based on the Lloyd-Max algorithm [10] in the federated learning setting. However, this method is computationally intensive and does not always yield an optimal solution. In summary, there lacks of a theoretical method to jointly optimize the truncation threshold and non-uniform quantizer parameters.

Our work aims to fill this gap. We innovatively combine gradient truncation with non-uniform quantization and extend this hybrid approach to a distributed learning framework. We address the joint optimization challenge and provide a novel solution that enhances the efficiency of communication without sacrificing the integrity of the learning process. The main contribution of this work is summarized as:

- We design a novel truncated non-uniform quantizer and integrate it into a distributed SGD framework under communication constraints.
- We provide a theoretical framework for analyzing the impact of the designed quantizer on convergence error.
- We derive optimal closed-form solutions for the with the assumption of Laplace gradient distribution.
- Both theoretical and numerical evaluations show that our proposed method outperforms the benchmarks and is even

competitive with the non-compressed models.

II. Problem Formulation

We consider a distributed learning problem, where N clients collaboratively train a shared model via a central server. The local dataset located at client i is denoted as $\mathcal{D}^{(i)}$. The objective is to minimize the empirical risk over the data held by all clients, i.e., solve the optimization problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} F(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i \mathbb{E}_{\xi^{(i)}\sim\mathcal{D}^{(i)}}\left[\ell(\boldsymbol{\theta};\xi^{(i)})\right], \tag{1}$$

where $w_i = \frac{|\mathcal{D}^{(i)}|}{\sum_{i=1}^{N}|\mathcal{D}^{(i)}|}$ is the weight of client i, $\xi^{(i)}$ is a random sample from $\mathcal{D}^{(i)}$ and $\ell(\boldsymbol{\theta};\xi^{(i)})$ is the local loss function of the model $\boldsymbol{\theta}$ at one data sample $\xi^{(i)}$. A standard approach to solve this problem is DSGD [2], [11], where each client i first downloads the global model $\boldsymbol{\theta}_t$ from the server at iteration t, then randomly selects a batch of samples $B_t^{(i)} \subseteq D^{(i)}$ of size B to compute its local stochastic gradient with respect to $\boldsymbol{\theta}_t$: $\boldsymbol{g}_t^{(i)} = \frac{1}{B}\sum_{\xi^{(i)}\in B_t^{(i)}} \nabla\ell(\boldsymbol{\theta}_t;\xi^{(i)})$. Then the server aggregates these gradients and updates the model: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\sum_{i=1}^{N} w_i \boldsymbol{g}_t^{(i)}$, where η is the server learning rate. We make the following two assumptions on the raw gradient $\nabla l(\boldsymbol{\theta}_t;\xi^{(i)})$ and the objective function $F(\boldsymbol{\theta})$ [12], [13]:

Assumption 1 (Bounded Variance)**.** *For parameter $\boldsymbol{\theta}_t$, the stochastic gradient $\nabla\ell(\boldsymbol{\theta}_t;\xi^{(i)})$ sampled from any local dataset have uniformly bounded variance for all clients:*

$$\mathbb{E}_{\xi^{(i)}\sim\mathcal{D}^{(i)}}\left[\|\nabla\ell(\boldsymbol{\theta}_t;\xi^{(i)}) - \nabla F(\boldsymbol{\theta}_t)\|^2\right] \le \sigma^2. \tag{2}$$

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

To reduce the communication cost, we compress the local stochastic gradients before sending them to the server: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\sum_{i=1}^{N} w_i \mathcal{C}_b[\boldsymbol{g}_t^{(i)}]$, where $\mathcal{C}_b[\cdot]$ is the compressor operator to compress each element of $\boldsymbol{g}_t^{(i)}$ into b bits. In this paper, we design the compressor using a two-stage quantizer.

III. Truncated Non-Uniform Quantizer for Distributed SGD

In this section, we introduce a two-stage quantizer that combines truncation with non-uniform quantization. Following this, we incorporate the quantizer into a distributed SGD algorithm and present an analysis of its convergence error.

Gradient Truncation The truncation operation cuts off the gradient so that the value is within a range. For an element g of gradient $\boldsymbol{g}$, the α-truncated operator $\mathcal{T}_\alpha[g]$ is defined as

$$\mathcal{T}_\alpha[g] = \begin{cases} g, & \text{for } |g| \le \alpha, \\ \text{sgn}(g)\cdot\alpha, & \text{for } |g| > \alpha \end{cases} \tag{3}$$

where α is a truncation threshold that determines the range of gradients, and $\text{sgn}(g)\in\{+1,-1\}$ is the sign of g. A common intuition is that the thicker the tail of the gradient distribution, the larger the value of α should be set to ensure that the discarded gradient information is upper bounded.

Nonuniform Gradient Quantization For the truncated gradient, we propose a novel element-wise non-uniform quantization scheme. Specifically, consider a truncated gradient element g that falls within the interval $[a_1, a_2]$. To satisfy communication constraints, we aim to encode it using b bits. This encoding process results in 2^b discrete quantization points, which effectively divide the interval $[a_1, a_2]$ into $s = 2^b - 1$ disjoint intervals. The boundaries of these intervals are defined by the points $a_1 = l_0 < l_1 \ldots < l_s = a_2$. Each k-th interval is denoted by $\Delta_k \triangleq [l_{k-1}, l_k]$, and has a length (or a quantization step size) of $|\Delta_k| = l_k - l_{k-1}$. If $g\in\Delta_k$, we have

$$\mathcal{Q}[g] = \begin{cases} l_{k-1}, & \text{with probability } 1-p_r, \\ l_k, & \text{with probability } p_r = \dfrac{g-l_{k-1}}{|\Delta_k|}. \end{cases} \tag{4}$$

It is evident that the specific operation of the quantizer depends on the quantization step size Δ_k, which is essentially the coded book $\mathcal{L} \triangleq \{l_0, l_1, ..., l_s\}$. This also determines the statistical characteristics of the quantizer, as demonstrated by the following lemma.

Lemma 1 (Unbiasness and Bounded Variance)**.** *For a truncated gradient element $g\in[a_1,a_2]$ with probability density function $p_g(\cdot)$, given the quantization points $\mathcal{L} = \{l_0, l_1, ..., l_s\}$, the nonuniform stochastic quantization satisfies:*

$$\mathbb{E}[\mathcal{Q}[g]] = g \tag{5}$$

and

$$\mathbb{E}\|\mathcal{Q}[g] - g\|^2 \le \sum_{k=1}^{s}\frac{P_k|\Delta_k|^2}{4} \tag{6}$$

where $P_k = \int_{l_{k-1}}^{l_k} p_g(x)\mathrm{d}x$ and $|\Delta_k| = l_k - l_{k-1}$.

The complete proof can be found in Appendix VII-A. We further introduce the concept of the "density" of quantization points, defined as $\lambda_s(g) \triangleq \frac{1}{|\Delta(g)|}$. This definition ensures that $\int_{a_1}^{a_2}\lambda_s(g)dg = s$. In the remainder of the paper, we denote a non-uniform quantizer with quantization destiny function$\lambda_s(\cdot)$ by $Q_{\lambda_s}[\cdot]$. By doing this, Lemma 1 can be rewritten as $\mathbb{E}[\mathcal{Q}_{\lambda_s}[g]] = g$ and $\mathbb{E}\|\mathcal{Q}_{\lambda_s}[g] - g\|^2 \le \int_{a_1}^{a_2}\frac{p(g)}{4\lambda_s(g)^2}\mathrm{d}g$. In a specific instance, if we take $\lambda_s(g) = \frac{s}{a_2-a_1}$, then the nonuniform quantization simplifies to uniform quantization [3], i.e., $\mathcal{L} = \{a_1 + k\frac{a_2-a_1}{s}, k = 0, 1, ..., s\}$.

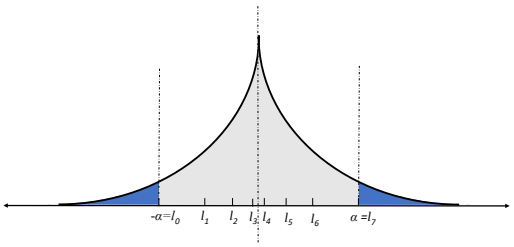

Fig. 1. Truncated Non-Uniform Quantizer (With truncation threshold $[-\alpha,\alpha]$ and quantization bit $b=3$ and quantization level $s=7$.)

To summarize, our proposed truncated non-uniform quan-

tizer, denoted as $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\boldsymbol{g})]$, begins with the truncation of gradients $\boldsymbol{g}$ using $\mathcal{T}_\alpha(\boldsymbol{g})$ to curtail values outside the $[-\alpha, \alpha]$ range, thereby reducing noise. These truncated gradients are then quantized through $\mathcal{Q}_{\lambda_s}[\cdot]$ into b-bit representations ($b = log_2(s+1)$), as depicted in Fig.1. The entire process is encapsulated in the Truncated Non-uniform Quantization for Distributed SGD (TNQSGD) algorithm, detailed in Algorithm1, which integrates our quantization method into the distributed SGD framework to enhance communication efficiency without compromising convergence performance.

Algorithm 1 Truncated Non-uniform Quantization for Distributed SGD (TNQSGD)

Input: Learning rate η, initial point $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, communication round T, parameters of $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\cdot)]$ (truncated threshold α, quantization density function λ_s);
for each communication rounds $t = 0, 1, ..., T-1$: **do**
  On each client $i = 1, ..., N$**:**
  Download $\boldsymbol{\theta}_t$ from server;
  Compute the local gradient $\boldsymbol{g}_t^{(i)}$ using SGD;
  Quantize $\boldsymbol{g}_t^{(i)}$ to $\hat{\boldsymbol{g}}_t^{(i)} = \mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\boldsymbol{g}_t^{(i)})]$ using Eq. (3) and (4);
  Send $\hat{\boldsymbol{g}}_t^{(i)}$ to the server;
  On the server:
  Aggregate all quantized gradients $\bar{\boldsymbol{g}}_t = \sum_{i=1}^N w_i \hat{\boldsymbol{g}}_t^{(i)}$;
  Update global model parameter: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \bar{\boldsymbol{g}}_t$;
end for

Assuming that each element follows a symmetrical probability density around zero $p(g)$ and is independently and identically distributed, we have the following Lemma to characterize the convergence performance of TNQSGD.

Lemma 2. *For a N-client distributed learning problem, by applying the quantizer $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\cdot)]$ and $w_i = \frac{1}{N}$, the convergence error of Alg. 1 for the smooth objective is upper bounded by*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\boldsymbol{\theta}_t)\|^2 \leq \underbrace{\frac{2[F(\boldsymbol{\theta}_0)-F(\boldsymbol{\theta}^*)]}{T\eta}+\frac{\sigma^2}{NB}}_{\triangleq \mathcal{E}_{DSGD}} + \underbrace{\frac{d}{4N}\int_{-\alpha}^{\alpha}\frac{p(g)}{\lambda_s(g)^2}\mathrm{d}g + \frac{2d}{N}\int_{\alpha}^{+\infty}(g-\alpha)^2 p(g)\mathrm{d}g}_{\triangleq \mathcal{E}_{TQ}} \quad (7)$$

Several insights can be derived from this Lemma. The term labeled $\mathcal{E}_{DSGD}$ in Eq. 7 represents the bound on the convergence error for the standard distributed SGD when applied with *non-compressed* model updates. The subsequent term, $\mathcal{E}_{TQ}$, encapsulates the error introduced by our proposed two-stage quantizer, which highlights the balance between the level of compression and the precision of our proposed algorithm. The term $\mathcal{E}_{TQ}$ can be further broken down into two components: a variance component due to quantization (the first term) and a bias component due to the truncation (the second term). It is important to note that with a sufficiently small truncation threshold α, the density of quantization points described by $\lambda_s(x)$ will be notably high. This concentration results in a reduction of the quantization variance towards zero, whilst simultaneously increasing the truncation bias. Conversely, as the threshold α increases, the truncation bias diminishes towards zero, but this leads to a rise in quantization variance. Additionally, the distribution of quantization points $\lambda_s(g)$ has a direct impact on the level of quantization variance, and thus influences the overall value of $\mathcal{E}_{TQ}$. For a detailed proof, refer to the Appendix VII-B.

IV. Optimal Parameters Design For Truncated Non-Uniform Quantizer

In this section, we aim to provide theoretical guidance for optimizing the parameters of the proposed quantizer. This can be addressed by solving a joint optimization problem involving two parameters, the truncation threshold and the quantization parameter.

A. Optimal Parameters for Any Gradient Distribution

Formally, we formulate the parameter selection problem as a *convergence error minimization* problem under the communication constraints:

$$\min_{\alpha,\lambda_s} \ \mathcal{E}_{TQ}(\alpha, \lambda_s) \quad s.t. \ \int_{-\alpha}^{\alpha}\lambda_s(x)dx = s \quad (8)$$

From Eq. (7), we find that only the first term of $\mathcal{E}_{TQ}$ contains the quantization density function $\lambda_s(g)$. Hence we investigate solution of $\lambda_s(g)$ by constructing the following Lagrange equation [14] using the variational principle:

$$I(\lambda_s(g), \nu) = \int_{-\alpha}^{\alpha}\Big[\frac{p(g)}{\lambda_s(g)^2} - \mu\lambda_s(g)\Big]\mathrm{d}g \quad (9)$$

To solve $\lambda_s(g)$ by applying the Euler-Lagrange equation:

$$-\frac{2p(g)}{\lambda_s(g)^3} - \mu = 0 \quad (10)$$

We can obtain $\lambda_s(g) = -(\frac{2p(g)}{\mu})^{\frac{1}{3}}$. Further using the communication budget constraints Eq. (8):

$$\lambda_s(g) = \frac{p(g)^{\frac{1}{3}}}{\int_{-\alpha}^{\alpha}p(g)^{\frac{1}{3}}\mathrm{d}g}\cdot s \quad (11)$$

From the above expression, we can derive some insights. Given fixed values of s and α, a larger $p(g)$ necessitates more quantization bits for effective compression. For a given gradient distribution $p(g)$ and communication constraint s, a larger truncation threshold α means retaining larger quantization range. As reflected in Eq. (11), this would increase the numerator, thereby decreasing the quantization points density.

Substitute Eq. (11) into Eq. (7), $\mathcal{E}_{TQ}$ can be rewrited as:

$$\mathcal{E}_{TQ}(\alpha) = \frac{d}{4Ns^2}\Big[\int_{-\alpha}^{\alpha}p(g)^{\frac{1}{3}}\mathrm{d}g\Big]^3 + \frac{2d}{N}\int_{\alpha}^{+\infty}(g-\alpha)^2 p(g)\mathrm{d}g \quad (12)$$

The optimum α can be found:

$$\alpha^* = \arg\min_{\alpha} \mathcal{E}_{TQ}(\alpha) \tag{13}$$

To examine the specific form of α, it is necessary to make an assumption about the $p(g)$, i.e., the distribution of the gradients. [4] demonstrates that the distribution of coordinates in each local gradient vector typically exhibits a bell-curve shape, akin to that of a Gaussian or Laplace distribution. In our work, we opt for the Laplace distribution which has heavier tails than the Gaussian distribution. This property allows us to capture the potentially larger deviations that are often present in the gradient values.

B. Optimal Parameters for Laplace Gradient Distribution

The Laplace gradient distribution is defined by:

$$p(g|\gamma) = Laplace(g|0,\gamma) = \frac{1}{2\gamma}\exp\{-\frac{|g|}{\gamma}\} \tag{14}$$

where γ is a scale parameter that indicates the range of variation in gradient values. A larger γ suggests a wider variation in gradient values, with heavier tails in the distribution.

Using the Laplace gradient distribution, Eq. (12) can be rewritten as:

$$\mathcal{E}_{TQ}(\alpha) = \frac{27d\gamma^2}{Ns^2}\{1-\exp[\frac{-\alpha}{3\gamma}]\}^3 + \frac{2d\gamma^2}{N}\exp[-\frac{\alpha}{\gamma}] \tag{15}$$

We can get α by minimizing Eq. (15):

$$\boxed{\alpha = 3\ln\left[1+\frac{\sqrt{6}s}{9}\right]\gamma} \tag{16}$$

The selection of quantization bit $b = 2,3,4$ (i.e., $s = 3,7,15$) leads to $\alpha = 1.79\gamma, 3.20\gamma, 4.88\gamma$, and $\mathcal{E}_{TQ} = \frac{0.61d\gamma^2}{N}, \frac{0.24d\gamma^2}{N}, \frac{0.077d\gamma^2}{N}$, respectively. Substituting Eq. (16) into Eq. (11) and using $p(g) = Laplace(g|0,\gamma)$, we can get

$$\boxed{\lambda_s(g) = \frac{3\sqrt{6}+2s}{8\gamma}\exp[\frac{-|g|}{3\gamma}]} \tag{17}$$

Eq. (16) and Eq. (17) represent the optimal solutions for the parameters of our designed two-stage quantizer. We next substitute them back into Eq. (7) to examine the minimum convergence error that can be guaranteed with this set of parameters. The result is shown in the following Theorem.

Theorem 1. *For a N-client distributed learning problem with constrained quantization level s, using α in Eq.* (16) *and $\lambda_s(g)$ in Eq.* (17)*, the convergence error of Alg. 1 for the smooth objective is upper bounded by*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\boldsymbol{\theta}_t)\|^2 \le \mathcal{E}_{DSGD} + \underbrace{\frac{27d\gamma^2}{N(s+\frac{3\sqrt{6}}{2})^2}}_{\mathcal{E}_{TQ}} \tag{18}$$

We can more intuitively observe how the communication constraint and gradient distribution affect model performance. First, there is a clear trade-off between the available communication resources (number of bits) and model convergence. To elaborate, when the communication constraint is more stringent, the number of bits available for communication is less (less s), which leads to a loss of information and poorer performance. Additionally, a larger γ suggests a wider variation in gradient values, with heavier tails in the distribution, which also hurts convergence. The complete proof can be found in Appendix VII-D.

C. Relationship With Other Quantization Scheme

We next examine the effects of truncation and quantization on gradient compression by considering three comparative scenarios based on whether truncation is applied and the type of quantization used—uniform or non-uniform. The scenarios are as follows:

Non-uniform Quantization without Truncation. In this case, we apply NSQ directly to the gradients without any preliminary truncation step, i.e., let $\alpha = \|\boldsymbol{g}\|_\infty \triangleq \max_j |g_j|$. We use the following lemma to character the upper bound of α, i.e., $\|\boldsymbol{g}\|_\infty$.

Lemma 3. *If the gradients follow Laplace distribution $Laplace(g|0,\gamma)$, $\|\boldsymbol{g}\|_\infty$ satisfied:*

$$\mathbb{E}[\|\boldsymbol{g}\|_\infty^2] \le 4\gamma^2[\ln 2d]^2 \tag{19}$$

The complete proof can be found in Appendix VII-C. Using Lemma 3 and Eq. (15), the error introduced by non-uniform quantization without truncation can be rewritten as:

$$\mathcal{E}_{TQ}^N \le \frac{27d\gamma^2}{Ns^2} \tag{20}$$

The selection of quantization bit $b = 2,3,4$ leads to $\mathcal{E}_{TQ}^N = \frac{3d\gamma^2}{N}, \frac{0.55d\gamma^2}{N}, \frac{0.12d\gamma^2}{N}$, respectively. By comparing equations Eq. (18) and Eq. (20), we can see that using only non-uniform quantization without truncation will incur a larger compression error.

Uniform Quantization with Truncation. Then, we apply truncation on gradients before being quantized uniformly. Truncation can help in reducing the range of gradient values to be quantized, which may lead to a more efficient uniform quantization process. Uniform quantization assigns equal-sized intervals for all values. That is, we set $\lambda_s(g) = \frac{s}{2\alpha}$, which is a typical uniform quantizer. Then $\mathcal{E}_{TQ}$ in Eq. (7) can be rewritten as:

$$\mathcal{E}_{TQ}^{UT}(\alpha) = \frac{d\alpha^2}{Ns^2} + \frac{2d\gamma^2}{N}\exp[-\frac{\alpha}{\gamma}] \tag{21}$$

Hence, we can obtain the optimal solution of α by minimizing Eq. (21):

$$\boxed{\alpha = v(s)\gamma} \tag{22}$$

where $v(s)$ satisfied $v\exp[v] = s^2$. The numerical result for bits $b = 2,3,4$ results with $\alpha = 1.68\gamma, 2.85\gamma, 4.02\gamma$, and $\mathcal{E}_{TQ}^{UT} = \frac{0.69d\gamma^2}{N}, \frac{0.28d\gamma^2}{N}, \frac{0.11d\gamma^2}{N}$, respectively. Substituting α in Eq. (22) to Eq. (21), we can derive the quantization error

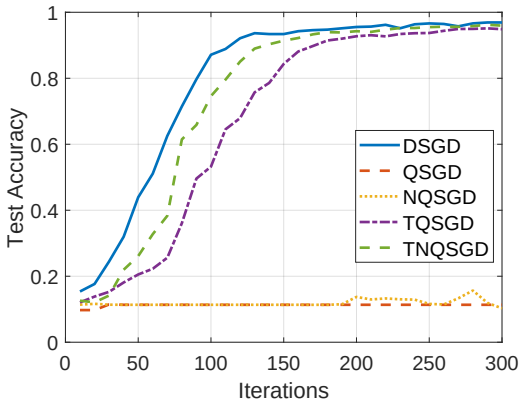

Fig. 2. Model Performance of Different Algorithms.

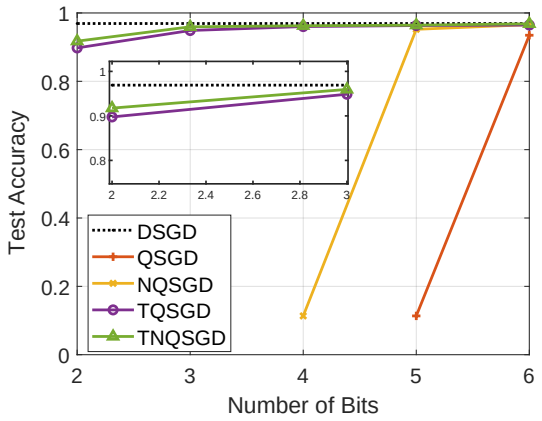

Fig. 3. Communication-Learning Tradeoff of Different Algorithms.

in this case as:

$$\mathcal{E}_{TQ}^{UT} = \frac{2d\gamma^2}{3Ns^2}[v(s)^2 + v(s)] \tag{23}$$

Compared to the results of truncation with non-uniform quantization (Theorem 1), we can see that we need to set a larger truncation threshold, which will result in a larger quantization error.

Uniform Quantization without Truncation. In this scenario, we apply uniform quantization to the full range of gradient values without any truncation. Using Lemma 3 and Eq. (21), the quantization error can be derived as:

$$\mathcal{E}_{TQ}^{U} \leq \frac{4d\gamma^2}{Ns^2}(\ln 2d)^2 \tag{24}$$

Take $d = 5 \times 10^5$ and $b = 2, 3, 4$, then $\mathcal{E}_{TQ}^{U} = \frac{84.83d\gamma^2}{N}, \frac{15.58d\gamma^2}{N}, \frac{3.39d\gamma^2}{N}$, respectively. We can see that using only uniform quantization without any truncation will result in errors far larger than the other three cases.

V. Experiments

In this section, we conduct experiments on MNIST to empirically validate our proposed TNQSGD. The MNIST consists of 70000 $1 \times 28 \times 28$ grayscale images in 10 classes. We compare our methods with three baselines discussed in previous section: 1) **QSGD** [3]: only use uniform quantization without truncation operation; 2) **NQSGD**: only use non-uniform quantization without truncation operation; 3) **TQSGD**: Combine Truncation and Uniform quantization; and 4) **DSGD**: as the oracle, clients send non-compressed gradients to the server.

Experimental Setting. We conduct experiments for $N = 8$ clients and use AlexNet [15] for all clients. We select the momentum SGD as an optimizer, where the learning rate is set to 0.01, the momentum is set to 0.9, and weight decay is set to 0.0005. Considered that gradients from convolutional layers and fully-connected layers have different distributions [4]. We thus quantize convolutional layers and fully-connected layers independently. We estimate γ based on maximum likelihood estimation: $\gamma = \frac{\sum_{j=1}^{d} |g_j|}{d}$.

Figure 2 illustrates the test accuracy of various algorithms on MNIST. DSGD achieves a test accuracy of 0.9691 with 32-bit full precision gradients. When $b = 3$ bits, TUQSGD and TNQSGD achieve test accuracies of 0.9487 and 0.9595, respectively. In contrast, QSGD and NQSGD are almost unable to converge. The results demonstrate that truncation operation can significantly improve the test accuracy of the model under the same communication constraints. Additionally, non-uniform quantization can further enhance the performance of the algorithm.

Figure 3 illustrates the tradeoff between communication budget and learning performance in terms of test accuracy of various algorithms. We compare this tradeoff between our proposed algorithms and two other baselines - QSGD and NQSGD. Additionally, we list the accuracy achieved by DSGD without communication budget constraints as a benchmark. All four algorithms exhibit a communication-learning tradeoff; that is, the higher the communication budget, the higher the test accuracy. However, our proposed TUQSGD and TNQSGD achieve higher test accuracies than the other two under the same communication cost.

VI. Conclusion

In conclusion, our work presents a two-stage quantization strategy that enhances communication efficiency in distributed SGD. The strategy incorporates an initial truncation step to reduce the influence of long-tail noise, succeeded by a non-uniform quantization tailored to the statistical properties of the truncated gradients. We conduct a thorough convergence analysis of the proposed algorithm. By optimizing the convergence error, we have formulated optimal closed-form solutions for setting the truncation threshold and non-uniform quantization parameters within specified communication constraints. Our findings and experimental results confirm that the proposed quantization approach surpasses existing methods, achieving superior convergence performance under equivalent communication constraints.

REFERENCES

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[2] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[3] 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.

[4] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] J. Chen, M. K. Ng, and D. Wang, "Quantizing heavy-tailed data in statistical estimation:(near) minimax rates, covariate quantization, and uniform recovery," *IEEE Transactions on Information Theory*, 2023.

[7] 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.

[8] V. Algazi, "Useful approximations to optimum quantization," *IEEE Transactions on Communication Technology*, vol. 14, no. 3, pp. 297–301, 1966.

[9] G. Chen, K. Xie, Y. Tu, T. Song, Y. Xu, J. Hu, and L. Xin, "Nqfl: Nonuniform quantization for communication efficient federated learning," *IEEE Communications Letters*, 2023.

[10] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[11] G. Yan, T. Li, S.-L. Huang, T. Lan, and L. Song, "Ac-sgd: Adaptively compressed sgd for communication-efficient distributed learning," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2678–2693, 2022.

[12] 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.

[13] D. Data and S. Diggavi, "Byzantine-resilient high-dimensional federated learning," *IEEE Transactions on Information Theory*, 2023.

[14] I. M. Gelfand, R. A. Silverman *et al.*, *Calculus of variations*. Courier Corporation, 2000.

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

VII. Appendix

A. Proof of Lemma 1

If $x \in \Delta_k$, we have

$$\mathcal{Q}[x] = \begin{cases} l_{k-1}, & \text{with probability } 1 - p_r, \\ l_k, & \text{with probability } p_r = \dfrac{x - l_{k-1}}{|\Delta_k|}. \end{cases}$$

Hence,

$$\begin{aligned} &\mathbb{E}\|\mathcal{Q}[x] - x\|^2 \\ &= \sum_{k=1}^{s} \int_{l_{k-1}}^{l_k} \Big[(l_{k-1} - x)^2 \frac{l_k - x}{|\Delta_k|} + (l_k - x)^2 \frac{x - l_{k-1}}{|\Delta_k|}\Big] p(x) \mathrm{d}x \\ &= \sum_{k=1}^{s} |\Delta_k|^2 \int_{l_{k-1}}^{l_k} \Big[\frac{l_k - x}{|\Delta_k|} \cdot \frac{x - l_{k-1}}{|\Delta_k|}\Big] p(x) \mathrm{d}x \\ &\overset{(b)}{\leq} \sum_{k=1}^{s} |\Delta_k|^2 \int_{l_{k-1}}^{l_k} \frac{p(x)}{4} \mathrm{d}x \\ &= \sum_{k=1}^{s} P_k \frac{|\Delta_k|^2}{4} \end{aligned}$$

where (b) uses $y(1-y) \leq \frac{1}{4}$ for all $y \in [0,1]$, and $P_k = \int_{l_{k-1}}^{l_k} p(x)\mathrm{d}x$.

B. Proof of Lemma 2

Firstly, we can decompose the mean squared error of the compressed gradient $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\boldsymbol{g})]$ into a variance term (due to the nonuniform quantization) and a bias term (due to the truncated operation):

$$\begin{aligned} \mathbb{E}[\|\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\boldsymbol{g})] - \boldsymbol{g}\|^2] = &\underbrace{d \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} \mathrm{d}g}_{\text{Quantization Variance}} \\ &+ \underbrace{2d \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) \mathrm{d}g}_{\text{Truncation Bias}}, \end{aligned} \tag{25}$$

Using the Assumption 1 and Eq. (25), we have

$$\begin{aligned} &\mathbb{E}[\|\bar{\boldsymbol{g}}_t - \nabla F(\boldsymbol{\theta}_t)\|^2] \\ &= \frac{\sigma^2}{BN} + \frac{d}{N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} \mathrm{d}g + \frac{2d}{N} \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) \mathrm{d}g \end{aligned} \tag{26}$$

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

$$F(\boldsymbol{\theta}') \leq F(\boldsymbol{\theta}) + \nabla F(\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{\theta}' - \boldsymbol{\theta}) + \frac{\nu}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2. \tag{27}$$

Hence, we can get

$$\begin{aligned} F(\boldsymbol{\theta}_{t+1}) &\leq F(\boldsymbol{\theta}_t) + \nabla F(\boldsymbol{\theta}_t)^{\mathrm{T}} (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{\nu}{2} \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2 \\ &= F(\boldsymbol{\theta}_t) - \eta \nabla F(\boldsymbol{\theta}_t)^{\top} \bar{\boldsymbol{g}}_t + \frac{\nu \eta^2}{2} \|\bar{\boldsymbol{g}}_t\|^2 \\ &\overset{(a)}{\leq} F(\boldsymbol{\theta}_k) - \eta \nabla F(\boldsymbol{\theta}_t)^{\top} \bar{\boldsymbol{g}}_t + \frac{\eta}{2} \|\bar{\boldsymbol{g}}_t\|^2 \\ &= F(\boldsymbol{\theta}_t) - \frac{\eta}{2} \|\nabla F(\boldsymbol{\theta}_t)\|^2 + \frac{\eta}{2} \|\bar{\boldsymbol{g}}_t - \nabla F(\boldsymbol{\theta}_t)\|^2 \end{aligned}$$

where (a) using $\eta \leq \frac{1}{\nu}$. Then using Eq. (26), we have

$$\begin{aligned} \mathbb{E}F(\boldsymbol{\theta}_{t+1}) \leq F(\boldsymbol{\theta}_t) &- \frac{\eta}{2} \|\nabla F(\boldsymbol{\theta}_t)\|^2 + \frac{\eta}{2NB} \sigma^2 \\ &+ \frac{d\eta}{2N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} \mathrm{d}g + \frac{\eta d}{N} \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) \mathrm{d}g \end{aligned}$$

Applying it recursively, this yields:

$$\begin{aligned} \mathbb{E}[F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_0)] \leq &-\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{\theta}_t)\|^2 + \frac{T\eta}{2NB} \sigma^2 \\ &+ \frac{dT\eta}{2N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} \mathrm{d}g + \frac{T\eta d}{N} \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) \mathrm{d}g \end{aligned}$$

Considering that $F(\boldsymbol{\theta}_T) \geq F(\boldsymbol{\theta}^*)$, so:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{\theta}_t)\|^2 \leq &\frac{2[F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)]}{T\eta} + \frac{\sigma^2}{NB} \\ &+ \frac{d}{N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} \mathrm{d}g + \frac{2d}{N} \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) \mathrm{d}g \end{aligned} \tag{28}$$

C. Proof of Lemma 3

The bound for $\|\boldsymbol{g}\|_\infty^2$ is attained by applying Markov's inequality to $f(\|\boldsymbol{g}\|_\infty^2) = \exp[\sqrt{\lambda \|\boldsymbol{g}\|_\infty^2}]$. For an arbitrary $\lambda > 0$,

$$\exp\{\sqrt{\lambda \mathbb{E}[\|\boldsymbol{g}\|_\infty^2]}\} \overset{(a)}{\leq} \mathbb{E}[\exp\{\sqrt{\lambda \max_j g_j^2}\}] \leq \sum_{j=1}^{d} \mathbb{E}[\exp\{\sqrt{\lambda}|g_j|\}]$$

where (a) follows from Jensen's inequality and definition of $\|\cdot\|_\infty$. Since $p(g|\gamma) = Laplace(g|0,\gamma)$,

$$\mathbb{E}[\exp\{\sqrt{\lambda}|g_j|\}] = \frac{1}{1 - \gamma\sqrt{\lambda}}$$

Therefore,

$$\mathbb{E}[\|\boldsymbol{g}\|_\infty^2] \leq \frac{1}{\lambda} \ln \frac{d}{1 - \gamma\sqrt{\lambda}}$$

Setting $\sqrt{\lambda} = \frac{1}{2\gamma}$ gives the desired bound in Lemma3.

D. Proof of Theorem 1

If we set $\alpha = 3\ln\left[1+\frac{\sqrt{6}s}{9}\right]\gamma$, then the truncated quantization error is

$$\mathcal{E}_{TQ} = \frac{27d\gamma^2}{N(s+\frac{3\sqrt{6}}{2})^2}$$

Replacing $\frac{d}{N}\int_{-\alpha}^{\alpha}\frac{p(g)}{4\lambda_s(g)^2}\mathrm{d}g + \frac{2d}{N}\int_{\alpha}^{+\infty}(g-\alpha)^2 p(g)\mathrm{d}g$ with $\mathcal{E}_{TQ}$ in Eq. (28), then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\boldsymbol{\theta}_t)\|^2 \leq \mathcal{E}_{DSGD} + \frac{27d\gamma^2}{N(s+\frac{3\sqrt{6}}{2})^2}$$