

Truncated Quantization for Heavy-tailed Gradients in Distributed Learning

Guangfeng Yan^a, Tan Li^b, Yuanzhang Xiao^c, Hanxu Hou^d and Linqi Song^a

^aDepartment of Computer Science, City University of Hong Kong, Hong Kong SAR

^bDepartment of Computer Science, The Hang Seng University of Hong Kong, Hong Kong SAR

^cDepartment of Electrical and Computer Engineering, University of Hawaii at Manoa, United States

^dSchool of Engineering and Intelligentization, Dongguan University of Technology, Dongguan, China

Abstract—Gradient compression has surfaced as a key technique to address the challenge of communication efficiency in distributed learning. In deep learning, however, it is observed that gradient distributions are heavy-tailed, with outliers significantly influencing the design of compression strategies. In this paper, we introduce a novel compression scheme specifically engineered for heavy-tailed gradients, which effectively combines gradient truncation with quantization. This scheme is adeptly implemented within a communication-limited distributed Stochastic Gradient Descent (SGD) framework. By assuming that the gradient tails follow a power-law distribution, we aim to minimize the error resulting from quantization, thereby determining optimal values for two critical parameters: the truncation threshold and the quantization density. We provide a theoretical analysis on the convergence error bound under both uniform and non-uniform quantization scenarios. Comparative experiments with other benchmarks demonstrate the effectiveness of our proposed method in managing the heavy-tailed gradients in a distributed learning environment.

Index Terms—Distributed Learning, Communication Efficiency, Heavy-tail Gradient, Power-law Distribution

I. INTRODUCTION

Distributed learning systems, which enable collaborative model training across multiple nodes or devices, have revolutionized the machine learning landscape. However, one of the most significant hurdles these systems face is the communication overhead. In distributed Stochastic Gradient Descent (DSGD) [1], [2], a widely adopted algorithm for distributed training, each communication round requires local clients to upload their model parameters to a central server for integration. With the increasing complexity of models and the growth in parameter size, this communication process has become a substantial bottleneck, challenging the practical limits of network bandwidth and efficiency.

To combat this issue, a variety of compression schemes have been introduced. Techniques such as sparsification [3], sketching [4], and quantization [5] have been explored to reduce the size of the transmitted data. Among these, quantization has gained widespread popularity due to its direct impact on reducing the number of bits required per parameter, making it a strategic fit for environments with limited communication resources. Employing low-bit representations [6], such as 2, 3, or 4 bits for the quantization of model gradients, presents a promising avenue for mitigating the communication load.

The effectiveness of gradient quantization techniques often hinges on the assumptions about the statistical distribution of gradients. Previous research efforts [5], [6] have designed quantization schemes based on assumptions that gradients follow Laplace or Gaussian distributions. However, our empirical analysis of gradients from real-world deep learning models reveals a distinctly heavy-tailed distribution, a crucial detail that has been largely overlooked. Fig. 1 clearly shows that both Laplace and Gaussian distributions exhibit tails that are too ‘thin’ to accurately estimate the true gradient distribution, which is more accurately modeled with ‘heavier’ tails.

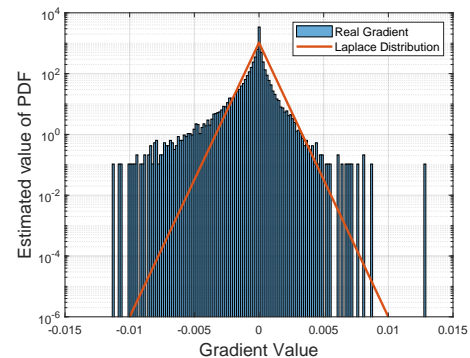


Fig. 1. The probability density of gradient computed with LeNet on MNIST. (The variance of the Laplace distribution is taken as the same value as the gradient variance.)

Researchers have adopted the Weibull distribution [7] as a more fitting representation of gradient distributions, leading to the design of non-uniform quantization schemes. However, these schemes have not been integrated with truncation techniques, which have been extensively used in communications [8]. Gradient truncation is essential for addressing the extreme values within a heavy-tailed distribution, which can significantly skew the quantization process. More importantly, existing quantization works have largely focused on single-client scenarios targeted at improving inference efficiency, rather than on the unique communication requirements of distributed learning environments.

In this paper, we introduce a novel quantization framework specifically engineered for distributed learning systems grappling with inherently heavy-tailed gradient distributions. Our

main contributions are as follows: (1) We propose a two-stage quantizer that initially truncates extreme gradient values before applying quantization, ensuring that communication overhead remains within a pre-determined budget. (2) We provide an in-depth analysis of how both truncation and quantization individually and jointly affect the quantization error terms in the convergence error bounds. (3) We assume that the tail of the gradient follows a power-law distribution, and determine the parameters of the design quantizer, namely, the truncation threshold and quantization density, by minimizing the quantization error. (4) We validate the effectiveness of our method through theoretical analysis of convergence performance and empirical experiments on actual deep learning 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

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

where $w_i = \frac{|\mathcal{D}^{(i)}|}{\sum_{i=1}^N |\mathcal{D}^{(i)}|}$ is the weight of client i , $\xi^{(i)}$ is random sampling from $\mathcal{D}^{(i)}$ and $\ell(\theta; \xi^{(i)})$ is the local loss function of the model θ towards one data sample $\xi^{(i)}$. A standard approach to solve this problem is DSGD [2], [9], where each client i first downloads the global model θ_t from server at iteration t , then randomly selects a batch of samples $B_t^{(i)} \subseteq \mathcal{D}^{(i)}$ with size B to compute its local stochastic gradient with model parameter θ_t : $\mathbf{g}_t^{(i)} = \frac{1}{B} \sum_{\xi^{(i)} \in B_t^{(i)}} \nabla \ell(\theta_t; \xi^{(i)})$. Then the server aggregates these gradients and updates the model: $\theta_{t+1} = \theta_t - \eta \sum_{i=1}^N w_i \mathbf{g}_t^{(i)}$, where η is the server learning rate. We make the following two common assumptions on the raw gradient $\nabla \ell(\theta_t; \xi^{(i)})$ and the objective function $F(\theta)$ [10], [11]:

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

$$\mathbb{E}_{\xi^{(i)} \sim \mathcal{D}^{(i)}} [\|\nabla \ell(\theta_t; \xi^{(i)}) - \nabla F(\theta_t)\|^2] \leq \sigma^2. \quad (2)$$

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

To reduce the communication cost, we consider to compress the local stochastic gradients $\mathbf{g}_t^{(i)}$ before sending them to the server:

$$\theta_{t+1} = \theta_t - \eta \sum_{i=1}^N w_i \mathcal{C}_b[\mathbf{g}_t^{(i)}] \quad (3)$$

where $\mathcal{C}_b[\cdot]$ is the operator to quantize each element of $\mathbf{g}_t^{(i)}$ into b bits. Next, we will introduce how to design this quantizer and determine its parameters.

III. TRUNCATED QUANTIZER FOR HEAVY-TAIL GRADIENTS

In this section, we introduce a two-stage quantizer designed to address the challenge of heavy-tailed gradients. We begin by detailing the two operative steps, truncation and general stochastic quantization, and analyze their impact on the convergence error.

A. Two-Stage Quantizer

In this paper, we form the quantizer using a two-stage operation.

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

$$\mathcal{T}_\alpha[g] = \begin{cases} g, & \text{for } |g| \leq \alpha, \\ \text{sgn}(g) \cdot \alpha, & \text{for } |g| > \alpha \end{cases} \quad (4)$$

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.

Gradient Quantization For the post-truncation gradient, we propose a general stochastic quantization scheme in an element-wise way. To clarify, 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 $a_1 = l_0 < l_1 < \dots < l_s = a_2$. 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

$$\mathcal{Q}[g] = \begin{cases} l_{k-1}, & \text{with probability } 1 - p_r, \\ l_k, & \text{with probability } p_r = \frac{g - l_{k-1}}{|\Delta_k|}. \end{cases} \quad (5)$$

It is evident that the specific operation of the quantizer depends on the quantization step size Δ_k , which is essentially the coded book $\mathcal{L} \triangleq \{l_0, l_1, \dots, l_s\}$. 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 $\mathcal{L} = \{l_0, l_1, \dots, l_s\}$, the nonuniform stochastic quantization satisfies:*

$$\mathbb{E}[\mathcal{Q}[g]] = g \quad (6)$$

and

$$\mathbb{E} \|\mathcal{Q}[g] - g\|^2 \leq \sum_{k=1}^s \frac{P_k |\Delta_k|^2}{4} \quad (7)$$

where $P_k = \int_{l_{k-1}}^{l_k} p_g(x) dx$ 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 density function $\lambda_s(\cdot)$ by $\mathcal{Q}_{\lambda_s}[\cdot]$. By doing this, Lemma 1 can be rewritten as $\mathbb{E}[\mathcal{Q}_{\lambda_s}[g]] = g$ and $\mathbb{E}\|\mathcal{Q}_{\lambda_s}[g] - g\|^2 \leq \int_{a_1}^{a_2} \frac{p(g)}{4\lambda_s(g)^2} dg$.

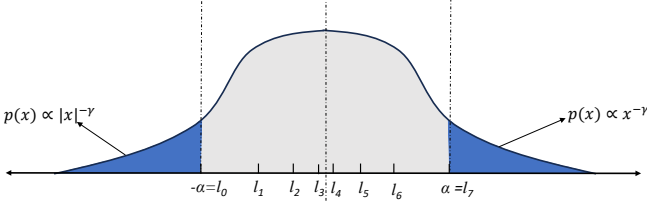


Fig. 2. Two-Stage Quantizer (With truncation threshold $[-\alpha, \alpha]$ and quantization bit $b = 3$ and quantization level $s = 7$.)

To summarize, our proposed two-stage quantizer, denoted as $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(g)]$, begins with the truncation of gradients g using $\mathcal{T}_\alpha(g)$ to curtail values outside the $[-\alpha, \alpha]$ range, thereby reducing the significant gradient noise. These truncated gradients are then quantized through $\mathcal{Q}_{\lambda_s}[\cdot]$ into b -bit representations. Please note that, as of now, we have not delineated a specific form for λ_s . One specific case emerges when $\lambda_s = \frac{s}{2\alpha}$, $\mathcal{Q}_{\lambda_s}[\cdot]$ becomes a uniform quantizer (as seen in the QSGD [5]), meaning that the truncated range of 2α , is evenly divided into s intervals. Conversely, when the intervals are not of equal size, $\mathcal{Q}_{\lambda_s}[\cdot]$ becomes a non-uniform quantizer.

For instance, in Fig. 2, we demonstrate a quantizer that integrates truncation with a non-uniform quantization density. Here, the interval $|l_4 - l_3| < |l_1 - l_0|$. This is due to a strategy that assign more quantization points to the peak of the distribution and fewer to the tails. The entire process is encapsulated in the Truncated Quantization for Distributed SGD (TQSGD) algorithm, detailed in Algorithm 1.

Algorithm 1 Truncated Quantizer for Distributed SGD (TQSGD)

- 1: **Input:** Learning rate η , initial point $\theta_0 \in \mathbb{R}^d$, communication round T , parameters of two-stage quantizer $\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\cdot)]$ (truncated threshold α , quantization density function λ_s);
 - 2: **for** each communication rounds $t = 0, 1, \dots, T - 1$: **do**
 - 3: **On each client** $i = 1, \dots, N$:
 - 4: Download θ_t from server;
 - 5: Compute the local gradient $g_t^{(i)}$ using SGD;
 - 6: Quantize $g_t^{(i)}$ to $\hat{g}_t^{(i)} = \mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(g_t^{(i)})]$ using Eq. (4) and (5);
 - 7: Send $\hat{g}_t^{(i)}$ to the server;
 - 8: **On the server:**
 - 9: Aggregate all quantized gradients $\bar{g}_t = \sum_{i=1}^N w_i \hat{g}_t^{(i)}$;
 - 10: Update global model parameter: $\theta_{t+1} = \theta_t - \eta \bar{g}_t$;
 - 11: **end for**
-

B. Performance Analysis

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 TQSGD.

Lemma 2. For a N -client distributed learning problem, by applying the two-stage 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

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 &\leq \underbrace{\frac{2[F(\theta_0) - F(\theta^*)]}{T\eta}}_{\triangleq \mathcal{E}_{DSGD}} + \frac{\sigma^2}{NB} \\ &+ \underbrace{\frac{d}{4N} \int_{-\alpha}^{\alpha} \frac{p(g)}{\lambda_s(g)^2} dg + \frac{2d}{N} \int_{\alpha}^{+\infty} (g - \alpha)^2 p(g) dg}_{\triangleq \mathcal{E}_{TQ}} \quad (8) \end{aligned}$$

Several insights can be derived from this Lemma. The term labeled \mathcal{E}_{DSGD} in Eq. 8 represents the bound on the convergence error for the standard distributed SGD when applied with *non-compressed* model updates. The second term, \mathcal{E}_{TQ} , indicates 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 attributed to the quantization process (the first term) and a bias component due to the truncation process (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 QUANTIZER PARAMETER DESIGN

In this section, we provide theoretical guidance for optimizing the parameters of the proposed quantizer. From the analysis of the process outlined in Alg. 1 and its performance, it is evident that determining the parameters of the two-stage quantizer, namely the truncation threshold α and the quantization density function $\lambda_s(g)$, is of critical importance. Formally, we formulate the parameter optimization problem as a *convergence error minimization* problem under the communication constraints:

$$\begin{aligned} \min_{\alpha, \lambda_s} \quad & \mathcal{E}_{TQ}(\alpha, \lambda_s) \\ \text{s.t.} \quad & \int_{-\alpha}^{\alpha} \lambda_s(x) dx = s \end{aligned} \quad (9)$$

A. Truncated Uniform Quantization

We first consider a simple case in which $\mathcal{Q}_{\lambda_s}[\cdot]$ operated as a uniform quantizer [5]. In this context, the quantization density is constant, $\lambda_s(x) = \frac{s}{2\alpha}$, leading to a coded book composed of evenly spaced points defined as $\mathcal{L} = \{a_1 + k\frac{2\alpha}{s}, k = 0, 1, \dots, s\}$. Consequently, our focus shifts to the optimization of α . To determine the optimal value of α , it is essential to assume the form of $p(g)$, the gradient distribution. In our analysis, we elect to adopt a widely recognized model - the power-law distribution [12].

Definition 1 (Power-law distribution [12]). A continuous power-law distribution is one described by a probability density $f(x|\gamma, x_{min})$ such that

$$f(x|\gamma, x_{min}) = \frac{\gamma-1}{x_{min}^{1-\gamma}} x^{-\gamma} \quad (10)$$

where x_{min} is the lower bound of Power-law distribution, γ is the tail index.

The density function of power-law distribution diverges as $x \rightarrow 0$, so Eq. (10) cannot hold for all $x \geq 0$; there exists a lower bound x_{min} to the power-law behavior. Hence, in this work, we only use power-law distribution to model the tails of gradients and ignore the intervals near 0, i.e., $\alpha > g_{min}$.

$$p(g|\gamma, g_{min}, \rho) = \rho \frac{\gamma-1}{g_{min}^{1-\gamma}} |g|^{-\gamma}, \quad \text{for } |g| > g_{min} \quad (11)$$

where $\rho = \int_{g_{min}}^{\infty} p(g)dg$, and $3 < \gamma \leq 5$. With the power-law distribution, the truncated quantization error in Eq. 8 can be rewrite as:

$$\mathcal{E}_{TQ}(\alpha) = \frac{dQ_U(\alpha)\alpha^2}{Ns^2} + \frac{4d\rho g_{min}^{\gamma-1}}{N(\gamma-2)(\gamma-3)} \alpha^{3-\gamma} \quad (12)$$

where $Q_U(\alpha) \triangleq \int_{-\alpha}^{\alpha} p(g)dg$. It is difficult to directly solve the above optimization problem. But through alternating iterations, we can obtain approximate numerical results:

$$\alpha = g_{min} \cdot \left[\frac{2\rho s^2}{(\gamma-2)Q_U(\alpha)} \right]^{\frac{1}{\gamma-1}} \quad (13)$$

The larger γ , the thinner the tail of gradients, and the larger the truncation parameter α . This aligns with our intuition. We use the following Theorem to characterize the convergence performance of Alg. 1 with the Truncated Uniform Quantizer (TQSGD).

Theorem 1. For an N -client distributed learning problem with quantization requirement s , the convergence error of Alg. 1 using $\lambda_s(g) = \frac{s}{2\alpha}$ and α in Eq. (13) for the smooth objective is upper bounded by

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 \leq \mathcal{E}_{DSGD} + (\gamma-1)Q_U(\alpha)^{\frac{\gamma-3}{\gamma-1}} \frac{dg_{min}^2(2\rho)^{\frac{2}{\gamma-1}} s^{\frac{6-2\gamma}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}} \quad (14)$$

Note that $Q_U(\alpha) \approx 1$. In practice, if we approximate α using $\alpha' \approx g_{min} \cdot \left[\frac{2\rho s^2}{(\gamma-2)} \right]^{\frac{1}{\gamma-1}}$, the convergence error becomes:

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 \leq \mathcal{E}_{DSGD} + [(\gamma-3)Q_U(\alpha') + 2] \frac{dg_{min}^2(2\rho)^{\frac{2}{\gamma-1}} s^{\frac{6-2\gamma}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}} \quad (15)$$

The difference between Eq.(14) and Eq.(15) lies in the coefficients of the second term. Specifically, the difference in their coefficients is $\epsilon = (\gamma-3)Q_U(\alpha') + 2 - (\gamma-1)Q_U(\alpha)^{\frac{\gamma-3}{\gamma-1}} \leq 2[1 - Q_U(\alpha')]$. Since $Q_U(\alpha')$ is close to 1, the value of ϵ is very small. This implies that the difference between the two equations is negligible.

B. Truncated Nonuniform Quantization

To be more general, we will next consider that the quantization density can be any positive function. Then the truncated quantization error in Eq. 8 can be rewrite as:

$$\mathcal{E}_{TQ}(\lambda_s(g), \alpha) = \frac{d}{4N} \int_{-\alpha}^{\alpha} \frac{p(g)}{\lambda_s(g)^2} dg + \frac{4d\rho g_{min}^{\gamma-1}}{N(\gamma-3)(\gamma-2)} \alpha^{3-\gamma} \quad (16)$$

The key question is how to jointly determine $\lambda_s(g)$ and α . We use the classic variational principle [13] to construct a Lagrange equation with only λ_s as a variable

$$I(\lambda_s(g), \nu) = \int_{-\alpha}^{\alpha} \left[\frac{p(g)}{\lambda_s(g)^2} - \mu \lambda_s(g) \right] dg \quad (17)$$

According to the Euler-Lagrange equation, we let

$$-\frac{2p(g)}{\lambda_s(g)^3} - \mu = 0 \quad (18)$$

Hence, $\lambda_s(g) = -(\frac{2p(g)}{\mu})^{\frac{1}{3}}$. Using the communication budget constraints Eq. (9), we further derive:

$$\lambda_s(g) = \frac{p(g)^{\frac{1}{3}}}{\int_{-\alpha}^{\alpha} p(g)^{\frac{1}{3}} dg} \cdot s \quad (19)$$

From the above expression, we can derive some insights. Given fixed values of s and α , a larger $p(g)$ necessitates more quantization points for effective compression. For a given gradient distribution $p(g)$ and communication constraint s , a larger truncation threshold α means retaining a larger quantization range. But unlike [14], [15], the above equation is integral to our analysis yet not in $\lambda_s(g)$'s conclusive form until we determine the truncation parameter α . We still use the assumption of power-law gradient tail in Eq. (11).

Let $Q_N(\alpha) \triangleq \int_{-\alpha}^{\alpha} p(g)^{\frac{1}{3}} (\frac{1}{2\alpha})^{\frac{2}{3}} dg$, then we can get the truncation threshold α through alternating iterations:

$$\alpha = g_{min} \cdot \left[\frac{2\rho s^2}{(\gamma-2)Q_N(\alpha)} \right]^{\frac{1}{\gamma-1}} \quad (20)$$

We use another Theorem to characterize the convergence performance of Alg. 1 with Truncated Non-Uniform Quantizer (TNQSGD).

Theorem 2. *For an N -client distributed learning problem with quantization requirement s , the convergence error of Alg. 1 using Eqs. (19) and (20) for the smooth objective is upper bounded by*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 &\leq \mathcal{E}_{DSGD} \\ &+ (\gamma - 1) Q_N(\alpha)^{\frac{\gamma-3}{\gamma-1}} \frac{dg_{min}^2 (2\rho)^{\frac{2}{\gamma-1}} s^{\frac{6-2\gamma}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}} \end{aligned} \quad (21)$$

Using the Holder's inequality, we can get:

$$\begin{aligned} Q_N(\alpha) &\triangleq \left[\int_{-\alpha}^{\alpha} p(g)^{\frac{1}{3}} \left(\frac{1}{2\alpha} \right)^{\frac{2}{3}} dg \right]^3 \\ &\leq \left[\left(\int_{-\alpha}^{\alpha} p(g) dg \right)^{\frac{1}{3}} * \left(\int_{-\alpha}^{\alpha} \frac{1}{2\alpha} dg \right)^{\frac{2}{3}} \right]^3 = Q_U(\alpha) \end{aligned}$$

This suggests that TNQSGD uses larger truncation threshold α , and achieve lower convergence error when compared to TQSGD.

V. EXPERIMENTS

In this section, we conduct experiments on MNIST to empirically validate our proposed TQSGD and TNQSGD methods. The MNIST consists of 70000 $1 \times 28 \times 28$ grayscale images in 10 classes. We compare our proposed methods with the following baselines: 1) **QSGD** [5]: only use uniform quantization without truncation ; 2) **NQSGD**: only use non-uniform quantization without truncation ; 3) **DSGD**: as the oracle, clients send non-compressed gradients to server.

Experimental Setting. We conduct experiments for $N = 8$ clients and use AlexNet [16] 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 [17]. We thus quantize convolutional layers and fully-connected layers independently. We estimate γ based on maximum likelihood estimation: $\gamma = 1 + n \left[\sum_{j=1}^n \ln \frac{g_j}{g_{min}} \right]^{-1}$, where $g_j, j = 1, \dots, n$, are the gradient values such that, $g_j > g_{min}$.

Figure 3 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, TQSGD and TNQSGD achieve test accuracies of 0.9515 and 0.9619, 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 4 illustrates the tradeoff between communication budget and learning performance in terms of test accuracy of various algorithms. We compare this tradeoff between our

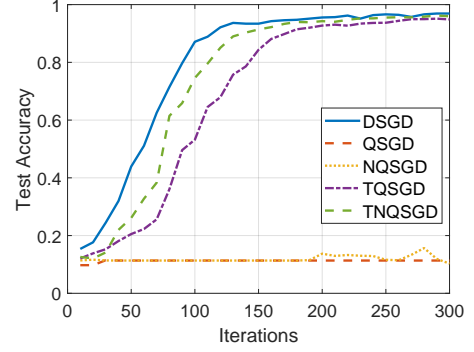


Fig. 3. Model Performance of Different Algorithms.

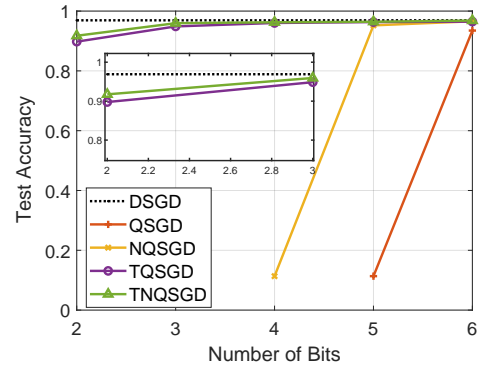


Fig. 4. Communication-Learning Tradeoff of Different Algorithms.

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 three algorithms exhibit a communication-learning tradeoff; that is, the higher the communication budget, the higher the test accuracy. However, our proposed TQSGD and TNQSGD achieve higher test accuracies than the other two under the same communication cost.

VI. CONCLUSION

We addressed the challenge of communication efficiency in distributed learning through a novel gradient quantization scheme designed for heavy-tailed distributions. Our approach, which bridges gradient truncation with quantization, is tailored for a communication-constrained distributed SGD environment. By assuming a power-law distribution for gradient tails, we refined the quantization process to minimize error and optimally set key parameters, truncation threshold and quantization density. Both our empirical results demonstrated that our method effectively manages heavy-tailed gradients, outperforming existing benchmarks in distributed settings.

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] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] Y. Liu, S. Rini, S. Salehkalaibar, and J. Chen, “M22: A communication-efficient algorithm for federated learning inspired by rate-distortion,” *arXiv preprint arXiv:2301.09269*, 2023.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] D. Data and S. Diggavi, “Byzantine-resilient high-dimensional federated learning,” *IEEE Transactions on Information Theory*, 2023.
- [12] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.
- [13] I. M. Gelfand, R. A. Silverman *et al.*, *Calculus of variations*. Courier Corporation, 2000.
- [14] 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.
- [15] V. Algazi, “Useful approximations to optimum quantization,” *IEEE Transactions on Communication Technology*, vol. 14, no. 3, pp. 297–301, 1966.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [17] 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.

VII. APPENDIX

A. Proof of Lemma 1

If $x \in \Delta_k$, we have

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

Hence,

$$\begin{aligned} & \mathbb{E} \|\mathcal{Q}[x] - x\|^2 \\ &= \sum_{k=1}^s \int_{l_{k-1}}^{l_k} \left[(l_{k-1} - x)^2 \frac{l_k - x}{|\Delta_k|} + (l_k - x)^2 \frac{x - l_{k-1}}{|\Delta_k|} \right] p(x) dx \\ &= \sum_{k=1}^s \int_{l_{k-1}}^{l_k} (x - l_{k-1})(l_k - x) p(x) dx \\ &\stackrel{(a)}{\approx} \sum_{k=1}^s \int_{l_{k-1}}^{l_k} (x - l_{k-1})(l_k - x) \frac{P_k}{|\Delta_k|} dx \\ &= \sum_{k=1}^s \frac{P_k}{|\Delta_k|} \frac{(l_k - l_{k-1})^3}{6} \\ &= \sum_{k=1}^s P_k \frac{|\Delta_k|^2}{6} \end{aligned}$$

where (a) uses the high-rate regime assumption, and $P_k = \int_{l_{k-1}}^{l_k} p(x) dx$.

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

where (b) uses $y(1 - y) \leq \frac{1}{4}$ for all $y \in [0, 1]$.

B. Proof of Lemma 2

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

$$\begin{aligned} \mathbb{E}[\|\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(\mathbf{g})] - \mathbf{g}\|^2] &= \underbrace{d \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} dg}_{\text{Quantization Variance}} \\ &\quad + \underbrace{2d \int_{\alpha}^{+\infty} (g - \alpha)^2 p(g) dg}_{\text{Truncation Bias}}, \end{aligned} \quad (22)$$

Using the Assumption 1 and Eq. (22), we have

$$\begin{aligned} & \mathbb{E}[\|\bar{\mathbf{g}}_t - \nabla F(\boldsymbol{\theta}_t)\|^2] \\ &= \frac{\sigma^2}{BN} + \frac{d}{N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} dg + \frac{2d}{N} \int_{\alpha}^{+\infty} (g - \alpha)^2 p(g) dg \end{aligned} \quad (23)$$

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

$$F(\boldsymbol{\theta}') \leq F(\boldsymbol{\theta}) + \nabla F(\boldsymbol{\theta})^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) + \frac{\nu}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2. \quad (24)$$

Hence, we can get

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

where (a) using $\eta \leq \frac{1}{\nu}$. Then using Eq. (23), we have

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

Applying it recursively, this yields:

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

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

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{\theta}_t)\|^2 &\leq \frac{2[F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)]}{T\eta} + \frac{\sigma^2}{NB} \\ &\quad + \frac{d}{N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} dg + \frac{2d}{N} \int_{\alpha}^{+\infty} (g - \alpha)^2 p(g) dg \end{aligned} \quad (25)$$

C. Proof of Theorem 1

If we take $\lambda_s(g) = \frac{s}{2\alpha}$ and use the assumption of Power-law Gradient Tail, then the truncated quantization error in Eq. 8

can be rewrite as:

$$\mathcal{E}_{TQ}(\alpha) = \frac{dQ_U(\alpha)\alpha^2}{Ns^2} + \frac{4d\rho g_{min}^{\gamma-1}}{N(\gamma-2)(\gamma-3)}\alpha^{3-\gamma}$$

Then we can the truncation threshold α by minimizing above equation:

$$\alpha = g_{min} \cdot \left[\frac{2\rho s^2}{(\gamma-2)Q_U(\alpha)} \right]^{\frac{1}{\gamma-1}}$$

Hence, the truncated quantization error is

$$\mathcal{E}_{TQ} = \frac{dg_{min}^2(\gamma-1)(2\rho)^{\frac{2}{\gamma-1}}Q_U(\alpha)^{\frac{\gamma-3}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}}s^{\frac{6-2\gamma}{\gamma-1}}$$

Replacing $\frac{d}{N} \int_{-\alpha}^{\alpha} \frac{p(g)}{4\lambda_s(g)^2} dg + \frac{2d}{N} \int_{\alpha}^{+\infty} (g-\alpha)^2 p(g) dg$ with \mathcal{E}_{TQ} in Eq. (25), then we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 \leq \frac{2[F(\theta_0) - F(\theta^*)]}{T\eta} + \frac{\sigma^2}{N} \frac{dg_{min}^2(\gamma-1)(2\rho)^{\frac{2}{\gamma-1}}Q_U(\alpha)^{\frac{\gamma-3}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}}s^{\frac{6-2\gamma}{\gamma-1}}$$

D. Truncated BiScaled Quantization

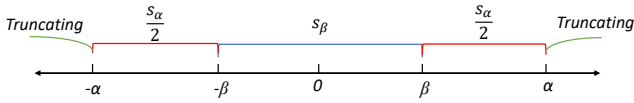


Fig. 5. Truncated BiScaled Quantization.

In this subsection, we consider:

$$\lambda_s(g) = \begin{cases} \frac{s_\beta}{2\beta}, & \text{for } g \in [-\beta, \beta], \\ \frac{s_\alpha}{2(\alpha-\beta)}, & \text{for } g \in [-\alpha, -\beta] \cup [\beta, \alpha], \end{cases} \quad (26)$$

where $s_\alpha + s_\beta = s = 2^b - 1$. And we still use the assumption of power-law gradient tail in Eq. (11). Let $p_1 \triangleq \frac{\int_0^\beta p(g)dg}{\beta}$ and $p_2 \triangleq \frac{\int_\beta^\alpha p(g)dg}{\alpha-\beta}$ denote the average probability density of g in $[0, \beta]$ and $[\beta, \alpha]$, respectively. Then

$$\mathbb{E}[\|\mathcal{Q}_{\lambda_s}[\mathcal{T}_\alpha(g)] - g\|^2] = \frac{2dp_1\beta^3}{s_\beta^2} + \frac{2dp_2(\alpha-\beta)^3}{s_\alpha^2} + \frac{4d\rho g_{min}^{\gamma-1}}{(\gamma-3)(\gamma-2)}\alpha^{3-\gamma}, \quad (27)$$

Then the truncated quantization error in Eq. 8 can be rewrite as:

$$\mathcal{E}_{TQ}(s_\alpha, s_\beta, \alpha, \beta) = \frac{2d\bar{p}_1\beta^3}{Ns_\beta^2} + \frac{2d\bar{p}_2(\alpha-\beta)^3}{Ns_\alpha^2} + \frac{4d\rho g_{min}^{\gamma-1}}{N(\gamma-3)(\gamma-2)}\alpha^{3-\gamma} \quad (28)$$

The key question is how to determine $s_\alpha, s_\beta, \alpha$ and β . To solve that, we formulate it as a convergence error minimization problem under the communication budget constraints:

$$\begin{aligned} \min_{s_\alpha, s_\beta, \alpha, \beta} \quad & \mathcal{E}_{TQ}(s_\alpha, s_\beta, \alpha, \beta) \\ s.t. \quad & s_\alpha + s_\beta = s, \end{aligned} \quad (29)$$

where $s = 2^b - 1$. By solving the above optimization problem, we firstly can determine s_α, s_β as:

$$s_\alpha = \frac{\bar{p}_2^{\frac{1}{3}}(1-k)}{\bar{p}_2^{\frac{1}{3}}(1-k) + \bar{p}_1^{\frac{1}{3}}k} \cdot s \quad (30)$$

$$s_\beta = \frac{\bar{p}_1^{\frac{1}{3}}k}{\bar{p}_2^{\frac{1}{3}}(1-k) + \bar{p}_1^{\frac{1}{3}}k} \cdot s \quad (31)$$

where $k \triangleq \frac{\beta}{\alpha}$. And let $Q_B(\alpha, k) \triangleq \left[(2 \int_{k\alpha}^\alpha p(g)dg)^{\frac{1}{3}}(1-k)^{\frac{2}{3}} + (2 \int_0^{k\alpha} p(g)dg)^{\frac{1}{3}}k^{\frac{2}{3}} \right]^3$, then

$$\mathcal{E}_{TQ}(\alpha, k) = \frac{dQ_B(\alpha, k)\alpha^2}{Ns^2} + \frac{4d\rho g_{min}^{\gamma-1}}{N(\gamma-3)(\gamma-2)}\alpha^{3-\gamma} \quad (32)$$

Hence, the optimum k and α can be found for TBQ as follows:

$$(k, \alpha) = \arg \min_k \mathcal{E}_{TQ}(\alpha, k) \quad (33)$$

Typically, there does not exist a closed solution for above minimization problem. A simple yet effective approximation is to use one step of alternating minimization between k and α : $k^* = \arg \min_k Q_B(\alpha, k)$, and

$$\alpha = g_{min} \cdot \left[\frac{2\rho s^2}{(\gamma-2)Q_B(\alpha, k^*)} \right]^{\frac{1}{\gamma-1}} \quad (34)$$

Using Eqs. (30) and (31), we can get the quantization point density function is

$$\lambda_s(g) = \begin{cases} \frac{\bar{p}_1^{\frac{1}{3}} \cdot s}{2\bar{p}_2^{\frac{1}{3}}(1-k^*)\alpha + 2\bar{p}_1^{\frac{1}{3}}k^*\alpha}, & \text{for } |g| \in [0, k^*\alpha], \\ \frac{\bar{p}_2^{\frac{1}{3}} \cdot s}{2\bar{p}_2^{\frac{1}{3}}(1-k^*)\alpha + 2\bar{p}_1^{\frac{1}{3}}k^*\alpha}, & \text{for } |g| \in [k^*\alpha, \alpha], \end{cases} \quad (35)$$

If we use Eqs. (35) and (34) to form our truncated quantization in Alg. 1, we can get Truncated BiScaled Quantization for Distributed SGD (TBQSGD). We characterize the convergence performance of TBQSGD in the following Theorem.

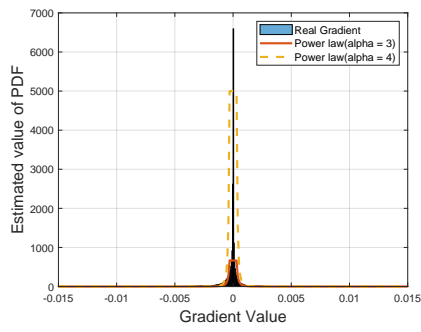
Theorem 3 (Convergence Performance of TBQSGD). *For an N -client distributed learning problem, the quantization bit is b , then the convergence error of TBQSGD for the smooth*

objective is upper bounded by

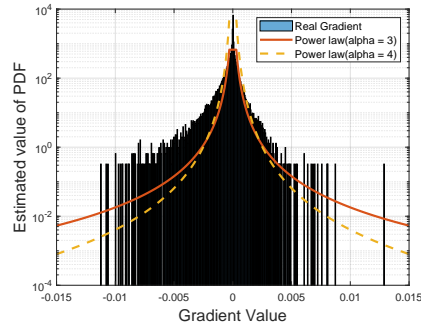
$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{\theta}_t)\|^2 &\leq \mathcal{E}_{DSGD} \\ &+ (\gamma - 1) Q_B(\alpha^*, k^*)^{\frac{\gamma-3}{\gamma-1}} \frac{dg_{min}^2(2\rho)^{\frac{2}{\gamma-1}} s^{\frac{6-2\gamma}{\gamma-1}}}{N(\gamma-3)(\gamma-2)^{\frac{2}{\gamma-1}}} \end{aligned} \quad (36)$$

Using the Holder's inequality, we can get $Q_B(\alpha^*, k^*) \leq 1$. This suggests that TBQSGD uses larger truncation threshold α , and achieve lower convergence error when compared to TUQSGD.

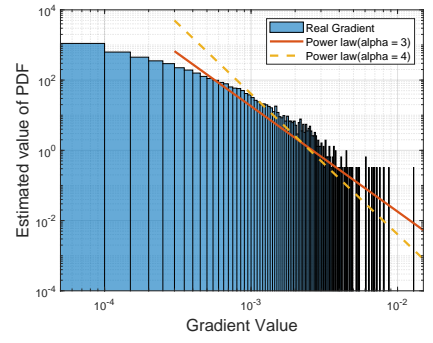
E. Experimental result



(a) Original



(b) Semilogy



(c) Loglog

Fig. 6. Estimating the probability density function of gradient computed with LeNet on MNIST. (Comparing the tail of gradient distribution and Power law distribution)